VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

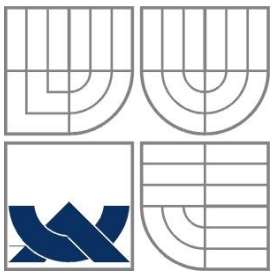
AUTOMATICKÁ TVORBA TEZAURU Z WIKIPEDIE

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

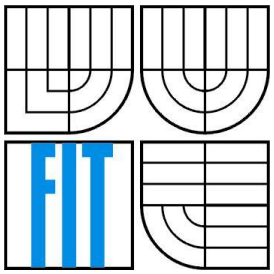
AUTOR PRÁCE
AUTHOR

Bc. Ján Novák

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ TVORBA TEZAURU Z WIKIPEDIE

ACQUIRING THESAURI FROM WIKIPEDIA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. Ján Novák
AUTHOR

VEDOUCÍ PRÁCE — Ing. Lubomír Otrusina
SUPERVISOR

BRNO 2011

Abstrakt

Tato práce se věnuje problematice automatické tvorby tezauru z Wikipedie. Obsahuje popis struktury Wikipedie jako vhodné datové sady pro tvorbu tezauru a popisuje některé metody výpočtu sémantické blízkosti termínů, které budou využity při tvorbě tezauru. Dále obsahuje popis návrhu a implementace systému pro automatickou tvorbu tezauru z Wikipedie. Na závěr je provedeno vyhodnocení výsledků systému.

Abstract

This thesis deals with automatic acquiring thesauri from Wikipedia. It describes Wikipedia as a suitable data set for thesauri acquiring and also methods for computing semantic similarity of terms are described. The thesis also contains a description of concepts and implementation of the system for automatic thesauri acquiring. Finally, the implemented system is evaluated by the standard metrics, such as precision or recall.

Klíčová slova

tezaurus, Wikipedie, Random Indexing, lexikální substituce, latentní sémantická analýza, sémantická blízkost termínů, sémantická podobnost termínů

Keywords

Thesauri, Wikipedia, Random Indexing, Lexical Substitution, Latent Semantic Analysis, Semantic Term Similarity, Semantic Term Relatedness

Citace

Novák Ján: Automatická tvorba tezauru z Wikipedie, diplomová práce, Brno, FIT VUT v Brně, 2011

Automatická tvorba tezauru z wikipedie

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením Ing. Lubomíra Otrusinu.
Další informace mi poskytl Ing. Marek Schmidt.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Ján Novák
25. května 2011

Poděkování

Za odborné vedení a cenné rady děkuji Ing. Lubomíru Otrusinovi a Ing. Marku Schmidtovi.

Obsah

1 Úvod

Tezaurus je množina prvkov a vzťahov medzi týmito prvkami. Prvky tezauru sú slová alebo frázy [12]. Medzi vzťahy patria synonymá, antonymá, hypernymá, hyponymá a iné. Jedno z hlavných využití tezauru v získavaní informácii je pri riešení problémov so slovnou zásobou, ktoré sa týkajú rozdielov medzi termínmi použitými v kolekcii dokumentov a termínmi, ktoré používa zadávateľ pri popise informácii, ktoré chce z tejto kolekcie získať [16]. Existuje veľa tezaurov, ktoré vytvorili ľudia, ako napríklad Rogetov alebo Macqurie.

Tvorba tezauru je veľmi časovo náročná úloha. Rovnako nie je ľahké udržovať ich aktuálne. Preto existuje snaha o ich automatickú tvorbu. Pri automatickej tvorbe tezauru sa na získavanie vzťahov medzi termínmi často používajú metódy výpočtu sémantickej blízkosti termínov. Hodnotenie blízkosti termínov je veľmi subjektívna úloha. Ako podobné sú si slová „auto“ a „lietadlo“? Obidve slová reprezentujú dopravné prostriedky. Je táto skutočnosť dostatočná na to, aby sme ich označili ako podobné? Pri tvorbe tezauru sa môžu použiť slovníky, ontológie a iné lexikálne štruktúry vytvorené človekom. Vytváranie takýchto štruktúr je však veľmi náročné, obzvlášť pri súčasnom prudkom rozvoji vedy a techniky a s tým súvisiacim rozširovaním slovnej zásoby. Preto je oveľa výhodnejšie, keď sú počítače schopné extrahovať tieto informácie samostatne, bez zásahu človeka. V takomto prípade sa veľmi často využívajú štatistické metódy, napríklad vektorové priestorové modely.

Automatické počítanie blízkosti termínov so sebou prináša celý rad problémov. Ľudská reč je vo svojej podstate veľmi nejednoznačná. Slová majú viac významov, ktoré často reprezentujú úplne rozdielne veci. Konštrukcie, ako je napríklad metafora, je človek schopný intuitívne rozlišovať. Počítač však túto schopnosť nemá. Aby schopný spracovávať prirodzený ľudský jazyk potrebuje analyzovať veľké množstvá textu.

Jedným zo zdrojov, ktoré obsahujú veľké množstvo textu je aj Wikipédia. Wikipédia je mnohojazyčná slobodná webová encyklopédia, na ktorej pracujú dobrovoľní prispievatelia z celého sveta [28]. Veľkosť Wikipédie a údaje v nej uložené sa využívajú aj v oblasti spracovania prirodzeného jazyka, kam patrí aj automatická tvorba tezauru.

Táto práca je založená na [9], kde je uvedená metóda na automatickú tvorbu tezauru z Wikipédie pomocou lexikálnej substitúcie. Prvá kapitola uvádza popis niektorých metód výpočtu sémantickej podobnosti termínov. V ďalšej kapitole je rozobraná štruktúra Wikipédie a je uvedený aj príklad, ako je možné túto štruktúru využiť v oblasti spracovania prirodzeného jazyka. Nasledujúce kapitoly obsahujú popis návrhu a implementácie systému pre automatickú tvorbu tezauru z Wikipédie a návrh testov, na ktorých sa bude vyhodnocovať kvalita systému. Posledná kapitola obsahuje vyhodnotenie výsledkov systému v uvedených testoch.

2 Metódy výpočtu sémantickej blízkosti termínov

V tejto kapitole budú prebrané niektoré metódy výpočtu sémantickej blízkosti termínov. Konkrétne sa bude jednať o latentnú sémantickú analýzu, Random Indexing a lexikálnu substitúciu. Všetky tieto metódy sú založené na distribučnej hypotéze. Distribučná hypotéza tvrdí, že podobné slová sa vyskytujú v podobných kontextoch [26]. Latentná sémantická analýza a Random Indexing sú metódy založené na slovnom priestorovom modeli. V tomto modele sú termíny reprezentovaná kontextovými vektormi, ktorých relatívne smernice určujú podobnosť termínov [27]. Na porovnanie kontextových vektorov sa môže použiť niekoľko metód. Najpoužívanejšie sú kosínová vzdialenosť, Euklidovská vzdialenosť, Manhattanská vzdialenosť.

Kosínová vzdialenosť určuje podobnosť vektorov na základe uhlu, ktorý zvierajú. Počíta sa zo skalárneho súčinu podľa vzorca:

$$\cos\varphi = \frac{\langle x, y\rangle}{\|x\|.\|y\|} \tag{2.1}$$

kde $\langle x, y\rangle$ znamená skalárny súčin vektorov x, y a $\|x\|$, resp. $\|y\|$, znamená veľkosť vektoru x, resp. y [5]. Kosínová vzdialenosť sa niekedy označuje aj ako normalizovaný skalárny súčin. Pred počítaním kosínovej vzdialenosti sa vektory často normalizujú (každý prvok vektoru sa vydelí veľkosťou vektoru). Tento krok zrýchľuje výpočet, pretože potom je možné kosínovu vzdialenosť vypočítať ako jednoduchý skalárny súčin.

Euklidovská vzdialenosť sa počíta podľa vzorca:

$$\|x - y\| = \sqrt{(x_1 - y_1)^2 + \ \ldots\ + (x_n - y_n)^2} \tag{2.2}$$

kde n je počet prvkov vektorov a x_m, resp. y_m, je m-tý prvok vektoru x, resp. y [5].

Manhattanská vzdialenosť sa tiež označuje ako taxicab vzdialenosť. Počíta sa podľa jednoduchého vzorca:

$$\|x - y\| = |x_1 - y_1| \ + \ \ldots\ + |x_n - y_n| \tag{2.3}$$

kde n je počet prvkov vektorov a x_m, resp. y_m, je m-tý prvok vektoru x, resp. y [5]. Manhattanská vzdialenosť je v podstate súčet absolútnych rozdielov jednotlivých prvkov vektoru.

Metódy založené na slovnom priestorovom modeli sa ukázali ako veľmi úspešne. V teste TOEFL (Test of English as a Foreign Language) dosiahla LSA úspešnosť 64,4% a Random Indexing 64,5 – 67 %. Po použití lematizácie, teda určení základných tvarov slov, čo znížilo počet nájdených jedinečných slov, stúpla úspešnosť Random Indexingu na 72%. Pre porovnanie zahraničný uchádzači na americké univerzity dosiahli priemernej úspešnosti 64,5% [27]. TOEFL je synonymický test,

v ktorom sú ku každému slovu ponúknuté štyri možnosti a uchádzač má vybrať to, ktoré najviac zodpovedá synonymu slova v otázke [13]. Pri náhodnom výbere odpovedí je teda očakávaní úspešnosť 25% [13]. Vyhodnocovanie metód prebiehalo tak, že kontextový vektor slova v otázke bol porovnaný s kontextovými vektormi ponúkaných slov a ako správna možnosť bol vybrané to slovo, ktorého vektor bol najpodobnejší vektoru slova v otázke [13].

2.1 Latentná sémantická analýza

Latentná sémantická analýza je teória a metóda na extrakciu a reprezentáciu významu slov v zmysle ich použitia v kontexte, pričom k tomu využíva štatistické výpočty aplikované na veľké textové korpusy [15]. Pri aplikácii tejto metódy v oblasti vyhľadávania informácii sa používa názov latentná sémantická indexácia.

Prvým krokom je vytvorenie matice výskytu termínov v kontextoch. Riadky matice reprezentujú termíny a stĺpce reprezentujú kontexty. Kontexty môžu byť napríklad dokumenty, odstavce alebo vety. Jednotlivé bunky matice obsahujú počet výskytov daného termínu v danom kontexte. Následne je táto matica transformovaná tak, že každá bunka je váhovaná funkciou. Táto funkcia by mala zabezpečiť, aby príliš často sa vyskytujúce termíny nemali príliš veľkú váhu. Rovnako je vhodné znížiť vplyv dlhých dokumentov na výsledok. Jednou z možností je funkcia, ktorá vyjadruje dôležitosť slova v danom kontexte a mieru informácie, ktorú toto slovo nesie v skúmanej doméne. Táto transformácia vyzerá tak, že ku každému prvku matice sa pripočíta hodnota 1. Následne sa spočíta entropia každého slova podľa vzorca $-p\ log\ p$ nad všetkými prvkami v danom riadku. Každá hodnota v riadku je vydelená hodnotou entropie pre daný riadok [15]. Riadky matice potom vytvoria vektory v mnohodimenzionálnom priestore tak, že prvky vektorov sú váhované frekvencie výskytu slov v jednotlivých kontextoch a dimenzionalita tohto vektorového priestoru je rovná počtu stĺpcov, teda počtu kontextov. Vektory nazývame kontextové vektory, lebo vyjadrujú kontext, v ktorom sa dané slovo vyskytuje. Keďže kontextové vektory reprezentujú distribučné profily slov, môžeme vyjadriť distribučnú podobnosť slov pomocou metód vektorovej podobnosti. Keďže reálne korpusy môžu obsahovať veľké množstvo dokumentov, vektory budú mať veľký počet dimenzií. Potom by bol výpočet podobnosti vektorov príliš časovo náročný. Matica výskytu termínov v kontextoch je obvykle riedka (väčšina hodnôt je nulová). Je to spôsobené tým, že len veľmi malá časť slov v jazyku sa vyskytuje vo veľkom počte kontextov. Ostatné sa vyskytujú len v malom počte. V bežnej matici výskytov je až 99 % prvkov nulových. Súvisí to so Zipfovým rozložením frekvencií výskytu slov v texte [27]. Zipfov zákon tvrdí, že frekvencia výskytu udalosti E, *P(E)*, v závislosti na jej ranku r je mocninná funkcia, ktorá sa dá vyjadriť vzťahom [29]:

$$P(E_r) \approx \frac{1}{r^{\alpha}} \tag{2.4}$$

Podľa tohto rozloženia sa najčastejšie vyskytujúce slovo objavuje v texte približne dvakrát častejšie ako druhé najfrekventovanejšie slovo, a to sa objavuje približne dvakrát častejšie ako štvrté a tak ďalej [18]. Napríklad v Brown Corpuse tvorí 135 slov polovicu obsahu. Na základe toho, že matica výskytov je riedka, redukuje sa počet dimenzií pomocou metódy singulárneho rozkladu (Singular Value Decomposition – SVD).

SVD má aj inú úlohu. S použitím pôvodných vektorov, ktoré mali dimenziu 30000, bola v teste TOEFL dosiahnutá úspešnosť 36%. S použitím redukovaných vektorov, ktoré boli získané metódou SVD a mali dimenziu 300, bola v tomto teste dosiahnutá úspešnosť 64%. Z toho bolo usúdené, že transformácia vektorov pomocou metódy SVD nejakým spôsobom korešponduje s ľudskou psychikou [13].

LSA je schopná zistiť blízkosť slov, aj keď sa spolu v texte vôbec neobjavia. Podľa tejto vlastnosti dostala táto metóda názov, pretože dokáže v texte odhaliť latentné vzťahy medzi termami a kontextami, teda vzťahy, ktoré nie sú viditeľné na prvý pohľad.

Hlavnou nevýhodou LSA je veľká náročnosť. Aj s použitím SVD trvá výpočet podobnosti termínov dlho, pretože samotné SVD je náročná operácia. Ďalšou nevýhodou je nemožnosť pridávať ďalšie kontexty. Pri pridaní ďalších kontextov je potrebné opäť počítať SVD. V poslednej dobe bolo vyvinutých niekoľko metód, ktoré umožňujú pridávať nové dokumenty bez toho, aby bolo nutné opätovne počítať singulárny rozklad pre už spracovanú množinu kontextov. Medzi tieto metódy patrí SVD-updating a Folding-in [14].

Folding-in je pomerne jednoduchá metóda. Nové dokumenty sú pred pridaním do matíc U a V premietnuté do priestoru redukovaných dokumentov respektíve termínov, čím sa premietne stav existujúcej databázy do týchto nových stĺpcových a riadkových vektorov. Stav existujúcej databázy sa však nepremietne do nových vektorov. Týmto vzniká vo výsledku chyba. Využitie metódy závisí najmä na počte zmien voči existujúcej databáze [14].

Na SVD-updating sa používa niekoľko metód, ktoré sú založené na rotáciách. Používajú sa napríklad ortonormálne μ-rotácie [10]. Ďalšou metódou je použitie QR-updatingu, ktorý využíva Givensove rotácie a následne aplikuje re-ortogonalizačnú metódu Jacobiho typu [22].

2.1.1 SVD

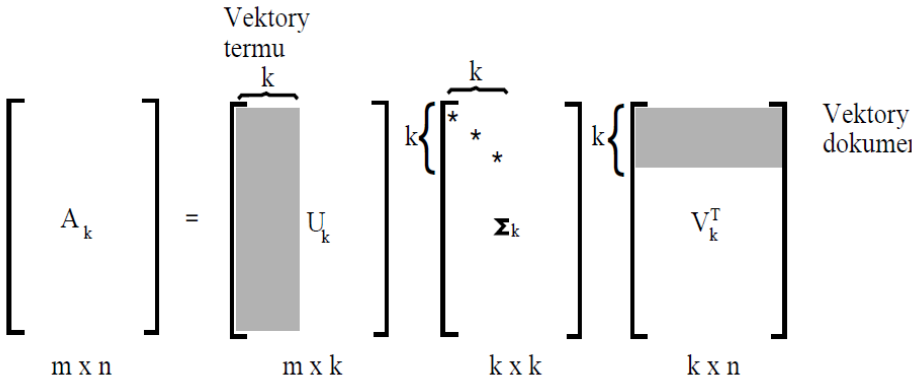
Singulárny rozklad je matematická metóda na rozklad matíc, ktorá je podobná faktorovej analýze. Jeho definícia je: „Nech A je ľubovoľná štvorcová matica. Potom existujú ortogonálne matice U a V a diagonálna matica Σ, na ktorej diagonále sú vlastné čísla matice $\sqrt{A^T A}$ tak, že $A = U\Sigma V^T$. Rozklad matice A na matice U, V a Σ sa nazýva singulárny rozklad matice A.“ [14]. Existuje dôkaz, ktorý tvrdí, že každú maticu je možné rozložiť s využitím najviac toľko faktorov, aká je najmenšia dimenzia originálnej matice [15].

Keďže matica termínov v dokumentoch obvykle nebýva štvorcová, ale býva rádu m×n, kde platí m≠n, matica U bude mať rozmery m×m, matica Σ m×n a matica V n×n. Rozklad pre m<n bude vyzerať nasledovne:

$$\begin{bmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{bmatrix} = \begin{bmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \sigma_m & \ldots & 0 \end{bmatrix} \begin{bmatrix} * & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & * \end{bmatrix} \tag{2.5}$$

$$\begin{array}{ccccc} A & = & U & & \Sigma & & V^T \\ m \times n & & m \times m & & m \times n & & n \times n \end{array}$$

Pri rozklade matice výskytu termínov v dokumentoch reprezentuje matica *U* kontextové vektory termínov, matica *V* reprezentuje maticu dokumentov. Σ je diagonálna matica m×n, ktorá obsahuje singulárne čísla $\sigma_1, \sigma_2, \ldots \sigma_{\min(m,n)}$. Tieto čísla sú usporiadané zostupne na hlavnej diagonále tak, že platí $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)}$. Toto zostupné usporiadanie znamená, že pre dostatočne presné výsledky nám stačí vypočítať prvých *k* najvyšších singulárnych čísel. Takto dostaneme k-aproximáciu hodnosti matice A. Vhodnú veľkosť čísla *k* treba určiť na základe experimentov. Pre veľké kolekcie dokumentov sa uvádza hodnota medzi 200 a 300 [14]. Takto sa redukuje mnohodimenzionálny priestor na priestor s dimenziou *k* a zároveň sa zachovávajú zhluky podobných dokumentov a termínov. K-aproximácia SVD je ukázaná na obrázku 2.1.



Obrázok 2.1 Znázornenie k-aproximácie singulárneho rozkladu

Výpočet SVD je veľmi náročný. Pri použití naivného algoritmu má faktoriálovú zložitosť [14]. V praxi sa však používajú numerické metódy, ktoré výpočet zrýchľujú. Rozklad matice m×n má zložitosť $O(\min(mn^2, m^2n))$ [11]. Pri LSA sa singulárny rozklad počíta iba pri indexácii a vyhľadávanie prebieha už vo vypočítanom rozklade. Boli vyvinuté techniky, ktoré zrýchľujú SVD. Niektoré využívajú aproximácie, napríklad pomocou metódy Monte Carlo [11]. V roku 2006 bola vypracovaná rýchla prírastková metóda, ktorá má pre r-aproximáciu matice veľkosti p×q časovú náročnosť O(pqr) , kde $r \leq \sqrt{min(p,q)}$ [3].

SVD je implementované napríklad v knižnici SVDLIBC, ktorá je založená na knižnici SVDPACKC [35]. Táto knižnica používa Lanczosovu a podpriestorovú metódu na určenie

singulárnych tripletov (singulárne hodnoty a zodpovedajúce pravé a ľavé vektory) [36]. Obdive metódy patria medzi iteratívne metódy.

2.2 Random Indexing

Random Indexing je slovný priestorový model, ktorý bol vytvorený ako alternatíva k LSA. Je založený na práci Pentti Kanevru o riedkych distribučných reprezentáciách [27]. Motivovaný je aj pozorovaním R. Hecht-Nielsena. Ten demonštroval, že v priestore s veľkým počtom dimenzií sa oveľa viac vyskytujú približne ortogonálne smery než tie skutočne ortogonálne. Z toho vyplýva, že môžeme použiť náhodné smery, aby sme vhodne aproximovali ortogonalitu [27]. Na základe tohto poznatku vzniklo niekoľko techník na redukciu počtu dimenzií. Najznámejšie sú Random Projection, Random Mapping a Random Indexing. Všetky tieto metódy sú založené na Johnson-Lindenstraussovej lemme. Tá hovorí, že keď premietneme body vektorového priestoru na náhodne vybraný podpriestor s dostatočne veľkým počtom dimenzií, vzdialenosti medzi bodmi budú približne zachované [27]. Preto môžeme počet dimenzií matice F redukovať jej násobením náhodnou maticou R:

$$F_{w\times d}R_{d\times k} = F'_{w\times k} \tag{2.6}$$

Dôležitým rozhodnutím je výber matice R. Pokiaľ by bola ortogonálna, bude platiť $F = F'$. Ak by bola približne ortogonálna, bude platiť $F \approx F'$. Najčastejšie sa používa Gaussovo rozloženie elementov náhodných vektorov v matici R. Avšak existuje aj jednoduchšia metóda. Skoro všetky elementy v týchto vektoroch budú nulové, čo znamená rozloženie s jednotkovou variáciou [27].

Hlavnou myšlienkou Random Indexing je akumulácia kontextových vektorov založených na výskyte slov v kontextoch. Každému kontextu je priradená unikátna reprezentácia, ktorá sa nazýva indexvektor. Indexvektor má dimenziu obvykle rádovo v tisícoch [27]. Je tvorený malým množstvom náhodne rozložených +1 a -1. Ostatné prvky sú 0. Vždy, keď sa slovo vyskytne v kontexte, je k jeho kontextovému vektoru pripočítaný indexvektor daného kontextu. To znamená, že slová sú reprezentované kontextovými vektormi, ktoré sú v podstate súčtom indexvektorov tých kontextov, v ktorých sa vyskytujú. Kontextami sú obvykle dokumenty alebo slová, môžu sa však využiť aj iné druhy kontextov.

Tento prístup je opačný než u LSA, kde najprv vytvoríme maticu spoluvýskytu a potom z nej extrahujeme kontextové vektory. Pri Random Indexingu najprv vytvoríme kontextové vektory a potom z nich môžeme zostaviť maticu spoluvýskytov tak, že použijeme kontextové vektory ako riadky matice. Takto vytvorená matica spoluvýskytov bude aproximáciou matice vytvorenej pomocou metódy LSA [27]. Metódou Random Indexingu však vznikne matica, ktorej dimenzionalita bude rádovo v tisíckach. Dosiahne sa teda to isté, na čo v LSA treba použiť SVD.

Metódou Random Indexing je možné vytvoriť aj klasickú maticu spoluvýskytov ako pri metóde LSA. Ak bude dimenzia indexvektorov rovná počtu kontextov a každý indexvektor bude obsahovať práve jednu 1 a ostatné hodnoty budú 0 a tieto vektory budú ortogonálne. Týmto postupom by sa však strácali výhody Random Indexingu.

Hlavnou výhodou Random Indexing je menšia náročnosť metódy. Nie je potrebné najprv tvoriť maticu spoluvýskytu, ale stačí hneď vytvoriť kontextové vektory. Odpadá potreba používať SVD na redukciu počtu dimenzií, čím sa šetrí pamäť aj čas.

Ďalšou výhodou je jednoduché pridávanie ďalších kontextov. Stačí vytvoriť nový indexvektor pre daný kontext a pripočítať ho ku kontextovým vektorom slov, ktoré sa vyskytujú v danom kontexte. Navyše pridanie kontextu nezvýši počet dimenzií. Ten sa pevne nastaví na začiatku ako parameter a neskôr sa už nemení.

2.3 Lexikálna substitúcia

Lexikálna substitúcia je textová následná podúloha, v ktorej systém poskytne niekoľko kandidátnych termínov *e* pre termín *w*, ktoré môžu byť dosadené do určitého kontextu $H_w = H^l w H^r$, pri čom sa vygeneruje kontext $H_e = H^l e H^r$, kde H^l označuje ľavý kontext a H^r označuje pravý kontext, z pôvodnej vety obsahujúcej slovo *w* [9]. Lexikálna substitúcia teda zoberie kontexty, v ktorých sa nachádza termín *w*, toto slovo v danom kontexte nahradí za termín *e*, o ktorom sa predpokladá, že sa jedná o podobný termín, a následne sa snaží vyhodnotiť kvalitu takto vzniknutej vety alebo jej časti.

Kvalita kontextu sa môže určiť podľa korpusu Web 1T 5-gram, ktorý obsahuje 1- až 5- slovné výrazy spolu s frekvenciou výskytov na Internete zistenou pomocou vyhľadávača Google [41]. Údaje boli získané z približne jedného bilióna slov na verejne prístupných webových stránkach. Kvôli veľkému množstvu rôznych n-gramov bolo potrebné ignorovať málo početné n-gramy. Preto sú v korpuse len n-gramy, ktorých frekvencia výskytu bola väčšia ako 40 [6]. Na vyhľadávanie kandidátnych termínov, ktoré budú do kontextov dosádzané, je možné použiť predpripravené slovníky alebo automatické metódy ako napríklad LSA alebo Random Indexing.

Pre výpočet podobnosti termínov je potrebné spraviť niekoľko krokov. Algoritmus na začiatku na vstupe očakáva pre každý termín množinu kontextov a množinu kandidátnych termínov. Následne sa vytvoria hypotetické frázy, v ktorých je daný termín nahradený kandidátnym termínom. Napríklad ak máme ako termín „book“ a kandidátny termín „novel“, tak z vety „The most printed book in history“ vygenerujeme hypotetickú frázu „The most printed novel in history“. Z týchto fráz sa vygenerujú n-gramy (1 < n <= 5) a ohodnotenie každého n-gramu sa získa ako hodnota pointwise mutual information (PMI) daného n-gramu vydelená hodnotou self-information (SI) ľavého a pravého kontextu. Tieto parametre sa rátajú podľa vzorcov:

$$PMI = log \frac{p(t_1, t_2 \dots t_n)}{p(t_1)p(t_2) \dots p(t_n)} \quad (2.7)$$

$$SI = -log\ p(t_1, t_2 \dots t_l) \quad (2.8)$$

kde $p(t_m)$ je frekvencia výskytu termínu t_m a $p(t_1, t_2, \dots t_m)$ je frekvencia výskytu fráze zloženej z termínov t_1 až t_m [9]. Na určovanie frekvencie výskytu kontextov a fráz sa používa Web 1T 5-gram korpus, ktorý obsahuje počet výskytov daného n-gramu na Internete. Frekvencia výskytu n-gramu sa získa ako podiel počtu výskytu tohto n-gramu a celkového počtu n-gramov rovnakej dĺžky vo Web 1T korpuse.

Delenie hodnotou self-information zaisťuje, že kontexty s nízkou informačnou hodnotou nebudú mať veľkú váhu [9]. Kontext s nízkou informačnou hodnotou je napríklad postupnosť stopslov (členy, zámená, spojky, ...).

Pre každý kandidátny termín bude celkové skóre rovné súčtu ohodnotení všetkých fráz, do ktorých bol tento kandidátny termín dosadený. Najlepšie kandidátne termíny budú mať najlepšie ohodnotenie.

3 Štruktúra Wikipédie

Wikipédia je na webe založená encyklopédia s otvoreným obsahom, ktorú možno voľne upravovať a slobodne čítať. Jednotlivé články sú upravované dobrovoľníkmi, takže články môže meniť hocikto. S tým súvisia problémy s nepresnosťou a vandalizmom.

Wikipédia má štruktúru hypertextového grafu, kde každý článok je prepojený s veľkým množstvom ostatných článkov pomocou hyperlinkov umiestnených priamo na termíny, ktoré vysvetľujú [20]. Tento graf je tak husto poprepájaný, že z článku sa dá dostať na hocijaký iný článok len pomocou priemerne 4,5 kliku [21]. Toto poskytuje veľa možností pre metódy vyhľadávania informácii. Veľkou prekážkou pre tieto metódy viacznačnosť termínov. V takomto prípade je odporúčaný postup za názov článku pridať do zátvoriek doplňujúcu informáciu o význame termínu. Nie vždy sa však prispievatelia držia tejto zásady. Niekedy sú vytvárane pseudočlánky, ktoré len presmerujú užívateľa na inú stránku. Na Wikipédii tiež existujú stránky, ktoré sú určené pre vyriešenie viacznačnosti termínov a obsahujú odkazy na články, ktoré sa zaoberajú rôznymi významami toho istého termínu.

3.1 Wikipedia Link Vector Model

Wikipedia Link Vector Model je technika, ktorá využíva hyperlinky medzi článkami Wikipédie na výpočet sémantickej blízkosti termínov [20]. Nevyužíva texty článkov, ale len ich názvy a odkazy. Podobnosť dvoch termínov sa vypočíta ako uhol medzi vektormi, ktoré reprezentujú daný článok. Tento vektor bude pozostávať zo všetkých odkazov v danom článku. Pre článok x, ktorý obsahuje odkazy na články $l_1 \dots l_n$, bude vektor vyzerať nasledovne:

$$x = (w(x \to l_1), w(x \to l_2), \dots, w(x \to l_n)) \tag{3.1}$$

kde $w(x \to l_m)$ znamená váha odkazu z článku x na článok l_m [20]. Váha odkazu z článku a na článok b sa vypočíta podľa nasledujúceho vzorca:

$$w(a \to b) = \left|a \to b\right| \log \left(\sum_{x=1}^{t} \frac{t}{\left|x \to b\right|}\right) \tag{3.2}$$

kde $\left|a \to b\right|$ znamená počet odkazov v článku a, ktoré odkazujú na článok b a t reprezentuje celkový počet článkov vo Wikipédii [20]. Počet výskytov daného odkazu v článku je teda násobený prevrátenou hodnotou pravdepodobnosti, že nejaký článok obsahuje odkaz na článok b. To zabezpečuje, že odkazy na články, na ktoré smeruje veľa odkazov, budú mať menšiu váhu.

Podobnosť článkov je daná uhlom, ktorý zvierajú vektory, ktoré ich reprezentujú. Veľkosť uhla sa pohybuje od 0° , čo znamená, že články obsahujú rovnaké linky, po 90°, čo značí, že linky

v článkoch sú úplne rozdielne. Ak je termín viacznačný a popisuje ho viac článkov, podobnosť termínov je rovná najmenšiemu uhlu medzi článkami.

Pri spracovávaní odkazov sa s viacznačnosťou narába nasledovne:

- Články sa spracujú priamo.
- Pokiaľ odkaz mieri na pseudočlánok, ktorý len presmerováva na iný článok, použije sa tento článok namiesto pseudočlánku.
- Stránky pre riešenie viacznačnosti termínov sú spracovávané tak, že sa použijú všetky články, na ktoré odkazujú.

3.2 SW1 korpus

SW1 korpus[1] je sémanticky anotovaná snímka anglickej Wikipédie z dňa 4.11.2006. Obsahuje 1 490 688 článkov. Obsahuje texty článkov a sémantické značky. Je rozdelený do 3 000 súborov, pričom každý obsahuje asi 500 článkov. Každý súbor začína niekoľkými riadkami komentárov, ktoré začínajú znakom '#', nasledovanými názvom súboru a názvami jednotlivých stĺpcov. Každý článok začína značkou %%#DOC <documentID> a každá veta končí značkou %%#SEN <sentence_number>. Ostatné riadky obsahujú texty článkov rozdelené na slová. Každé slovo je na samostatnom riadku, ktorý je rozdelený na stĺpce, ktoré obsahujú dané slovo, jeho slovný druh, základný tvar a značky pre jednotlivé zdroje termínov. Zdroje termínov sú: CONL, WordNet, Wall Street Journal, odhad anafor a hyperlinky. Značky pre zdroje termínov sa delia na:

- B-značka – označuje začiatok termínu
- I-značka – označuje pokračovanie termínu
- 0-značka – žiadna značka

Termíny teda pozostávajú zo slova, ktoré má B-značku, pokračujú 0 až N slovami, ktoré majú I-značku a končia slovom, ktoré má B-značku, 0-značku alebo koncom vety. Niekedy sa môže vyskytnúť aj termín, ktorý nezačína B-značkou ale I-značkou [34].

Tento korpus poslúži ako prostriedok pre vyhľadávanie termínov vo Wikipédii, pretože sú v ňom označené termíny, základné tvary slov a slovné druhy. Všetky tieto údaje sú pri extrakcii termínov potrebné, a tým, že sú v tomto korpuse, sa ušetrí čas, ktorý by bol potrebný na predspracovanie Wikipédie.

[1] http://barcelona.research.yahoo.net/dokuwiki/doku.php?id=semantically_annotated_snapshot_of_wikipedia

4 Návrh systému

Táto kapitola sa zaoberá návrhom a architektúrou vyvíjaného systému. Vlastná implementácia je rozobraná v nasledujúcej kapitole. V nasledujúcom texte sú slovným spojením skúmané termíny označené termíny, ktoré sú systému zadané a pre ktoré má systém nájsť podobné termíny. Slovným spojením kandidátne termíny alebo slovom kandidáti sú označené termíny, ktoré systém vybral v prvej fázy ako najpodobnejšie pre jednotlivé skúmané termíny. Spojením potenciálne kandidátne termíny sú označované termíny, ktoré systém dostal na vstupe a z ktorých sa vyberajú jednotlivý kandidáti pre skúmané termíny. Pokiaľ by takýto zoznam nebol zadaný, sú za potenciálne kandidátne termíny považované všetky termíny v indexe. Rovnako pokiaľ by nebol zadaný zoznam skúmaných termínov, sú za skúmané termíny považované všetky termíny v indexe.

V tejto kapitole je najprv predstavená architektúra systému a prepojenie jednotlivých jeho častí. Každá časť je potom popísaná v samostatnej podkapitole.

4.1 Architektúra systému

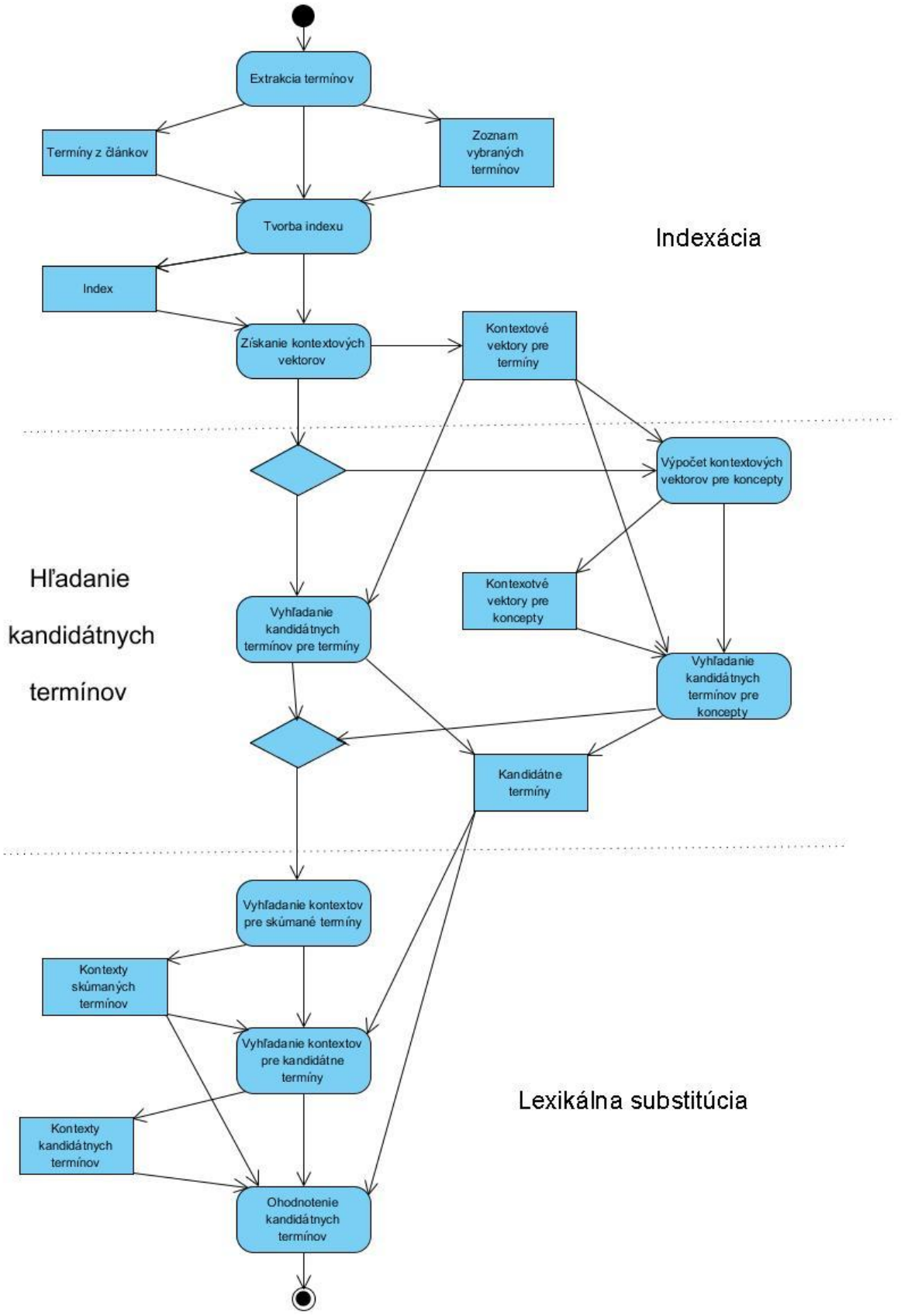
Celé spracovanie má tri fáze. Každá fáza je tvorená niekoľkými modulmi, ktoré zabezpečujú funkcionalitu systému. Celý systém má na vstupe SW1 korpus. Voliteľne môže mať na vstupe zoznam skúmaných a potenciálnych kandidátnych termínov.

Prvou fázou je tvorba indexu, ktorá pozostáva z extrakcie termínov, tvorby vlastného indexu a vytvorenia kontextových vektorov pre termíny.

Druhou fázou je vyhľadávanie kandidátnych termínov pomocou metódy Random Indexing. Táto fáza má dve varianty. Prvou je porovnávanie kontextových vektorov pre jednotlivé termíny. V tomto prípade sa porovnávajú termíny s termínmi. Druhým variantom je porovnávanie konceptov s termínmi. V tomto prípade je nutné najprv vytvoriť kontextové vektory pre koncepty a tie následne porovnať s kontextovými vektormi pre termíny. Výstupy z obidvoch variantov sú zhodné, takže výber variantu neovplyvní ďalší beh systému.

Treťou a poslednou fázou je lexikálna substitúcia, kde sa kandidátne termíny preusporiadajú pomocou lexikálnej substitúcie. Výstup tejto fáze je zároveň výstupom celého systému a obsahuje zoznam podobných termínov pre každý skúmaný termín. Tieto termíny sú usporiadané podľa podobnosti od najpodobnejšieho k najmenej podobnému.

Diagram 4.1 ukazuje jednotlivé moduly systému, vrátane ich následnosti. Zobrazuje tiež vstupy a výstupy jednotlivých modulov. Moduly sú ďalej v texte detailne popísané v takom poradí, v akom sú zobrazené v diagrame.



Obrázok 4.1 Architektúra systému

4.2 Filtrovanie termínov

V indexe je potrebné mať len kvalitné termíny, ktoré spĺňajú isté predpoklady. Ako termíny môžu byť totiž označené aj slová alebo slovné spojenia, ktoré v skutočnosti nie sú termínmi a do systému by mohli zanášať nepresnosti. Typickým príkladom sú čísla a číslice, ktoré sú aj v SW1 korpuse označené ako termíny, ale v skutočnosti termínmi nie sú. Nenesú ani žiadnu užitočnú informáciu o kontexte.

Na filtrovanie sa používa niekoľko metód. Prvou metódou je filtrovanie pomocou stoplistu. Slová, ktoré sa nachádzajú v stoplist̃e, sú zamietnuté. Použil sa bežný stoplist pre anglické texty[2], ktorý je doplnený o špecifické slová a frázy, ktoré sa vyskytujú na Wikipédii príliš často, ako napríklad „citation needed“, „reference“, „article“ a podobne. Rovnako sú do stoplistu pridané príliš všeobecné slová, ako napríklad príslovky, ktoré boli označené ako termíny a počas testovania sa objavili vo výsledkoch príliš často vybrané ako kandidátne termíny k nesúvisiacim skúmaným termínom. Medzi tieto slová a fráze patria napríklad „for the most part“, „as well“, „possibly“ a podobne.

Ďalšou metódou je filtrovanie na základe dĺžky slova a znakov v slove. Termíny musia mať aspoň tri znaky. Kratšie zmysluplné termíny sa prakticky nevyskytujú, okrem skratiek, ktoré môžeme zanedbať. Ďalej termín musí obsahovať viac písmen ako ostatných znakov. Pokiaľ má termín len tri znaky, musia to byť všetko písmená. Týmto sa odstránia napríklad anglické radové číslovky, ktoré majú síce tri znaky, ale jeden z nich je číslica. Pri radových číslach väčších ako 9 zas nie je splnená podmienka, že obsahujú viac písmen ako ostatných znakov (skladajú sa z dvoch číslic a dvoch písmen). Odstraňujú sa aj termíny, ktoré obsahujú iné ako povolené znaky. Povolené znaky sú písmená, číslice, a znaky pomlčka (-), bodka (.), apostrof (`) a podtržník (_).

Tento filter sa využíva v celom systéme, predovšetkým pri extrakcii termínov. Okrem toho sa používa napríklad pri extrakcii kontextov pre skúmané termíny alebo filtrovanie zadaných zoznamov skúmaných a potenciálnych kandidátnych termínov.

4.3 Extrakcia termínov

Termíny sa vyhľadávajú v SW1 korpuse. Použité sú termíny zo všetkých zdrojov SW1 korpusu a navyše sú extrahované aj slovesá, podstatné a prídavné mená. Všetky termíny sú extrahované v jednom prechode korpusom za pomoci značiek, ktoré SW1 korpus obsahuje. Výstup zo systému má dve časti.

Prvou časťou je zoznam termínov v jednotlivých článkoch Wikipédie. Pre každý článok sú uvedené všetky termíny. Každý termín bude uvedený toľkokrát, koľkokrát sa vyskytol v článku.

[2] http://www.thebananatree.org/stoplist.html

Zachované je aj poradie výskytov jednotlivých výskytov termínov. Výstup teda vyzerá ako článok Wikipédie, v ktorom boli odstránené ignorované slová a zostali iba termíny.

Ďalšou časťou výstupu je zoznam termínov, ktoré prešli filtrom. Vo Wikipédii sa totiž môžu vyskytnúť aj termíny, ktoré nespĺňajú podmienky na to, aby boli indexované. Ignorované sú všetky málo početné termíny a tiež termíny, ktoré neprešli filtrom uvedeným v kapitole 4.2.

4.4 Tvorba indexu

Index sa tvorí pomocou knižnice Apache Lucene[3]. Táto knižnica sa využíva preto, lebo sa jedná o vysokovýkonnú open-sourcovú knižnicu pre indexáciu a vyhľadávanie, takže v prípade potreby je možné ju jednoducho upraviť. Vstup indexeru je zhodný s výstupom z extrakcie termínov. Na vstupe je teda zoznam termínov v jednotlivých článkoch a zoznam skutočných termínov, ktoré prešli filtrom. Indexer postupne prechádza súbor, ktorý obsahuje termíny pre jednotlivé články, pričom sa ignorujú tie termíny, ktoré nie sú uvedené v zozname. Do indexu sa uloží názov článku a termíny z článku, ktoré sú uvedené v zozname termínov. Pri týchto termínoch zostane zachovaný počet aj poradie ich výskytov.

4.5 Získanie kontextových vektorov

Kontextové vektory sa získavajú metódou Random Indexing. Na ich výpočet je použitá knižnica semanticvectors[4]. Je to open-sourcová knižnica šírená pod BSD licenciou. Obsahuje implementáciu algoritmov Random Projecting a Random Indexing. Pracuje nad Apache Lucene indexom.

Pre každý termín v indexe je vytvorený vlastný kontextový vektor. Uložia sa dve verzie kontextových vektorov, a to nenormalizovaná a normalizovaná. Rovnako sa uložia aj náhodné vektory pre všetky termíny. V ďalšom priebehu sa počíta s normalizovanými vektormi. Nenormalizované kontextové a náhodné vektory sa ukladajú najmä kvôli možnosti rozšírenia indexu. V takom prípade je možné využiť už vypočítané časti a iba k nim pridať nové dokumenty. Odpadá tak potreba znovu spracovávať už spracované dokumenty. Náhodné vektory sa ukladajú aj kvôli výpočtu kontextových vektorov pre koncepty, keď sú dve možnosti výpočtu. Jedna možnosť počíta s normalizovanými kontextovými vektormi a druhá s náhodnými vektormi.

Normalizácia prebieha podľa vzorca:

$$v_i = \frac{v_i}{\|v\|} \tag{4.1}$$

kde v_i je i-tý prvok vektoru a ||v|| je veľkosť vektoru, ktorá sa vypočíta podľa vzorca:

[3] http://lucene.apache.org/java/docs/
[4] http://code.google.com/p/semanticvectors/

$$\|v\| = \sqrt{\sum_{i=0}^{n} v_i^2} \tag{4.2}$$

kde n je počet prvkov vektoru. Počet prvkov vektoru sa tiež nazýva dimenzia vektoru. Vďaka tejto úprave je možné kosínovú podobnosť (vzorec 2.1) nahradiť skalárnym súčinom, pričom zostane zachovaná výsledná hodnota. Toto výrazne zvýši rýchlosť výpočtu podobnosti vektorov, pretože veľkosť vektoru sa bude počítať iba jedenkrát, a to pri normalizácii. Pri kosínovej vzdialenosti by sa veľkosť vektoru počítala pri každom porovnávaní vektorov.

4.6 Tvorba kontextových vektorov pre koncepty

Pod pojmom koncepty sú chápané jednotlivé články na Wikipédii. Jednotlivé termíny môžu mať viac významov. Články na Wikipédii však súvisia iba s jednou témou. Preto sú ideálnymi kandidátmi na koncepty. Koncepty sú reprezentované termínmi, ktoré sa nachádzajú v príslušnom článku na Wikipédii. Do úvahy sa berú len tie termíny, ktoré sú v indexe. Neberú sa teda do úvahy termíny, ktoré neprešli filtrom pri vyhľadávaní. Termíny môžu byť reprezentované buď svojim kontextovým vektorom alebo svojim náhodným vektorom. Pri kontextových vektoroch sa používa normalizovaná forma, pretože v nenormalizovanej forme majú kontextové vektory pre veľmi početné termíny vysoké hodnoty a tieto vysoké hodnoty by ovplyvňovali výsledky. Tvorba kontextových vektorov pre koncepty má dve varianty podľa toho, aké vektory pre termíny sa použijú.

Kontextový vektor pre koncept bude súčtom vektorov termínov konceptu. Každý termín bude pre každý koncept váhovaný podľa toho, ako špecifický je pre tento koncept a zároveň podľa toho, ako veľmi je daný termín významný pre tento koncept v porovnaní s ostatnými termínmi konceptu. Vypočíta sa váha termínu a touto váhou sa vynásobí jeho vektor predtým, ako sa pripočíta ku kontextovému vektoru konceptu. Váhovanie prebieha pomocou metódy tf-idf. Váha termínu *i* v dokumente *j* sa vypočíta podľa vzorca:

$$tfidf_{i,j} = tf_{i,j} * \log\frac{|D|}{df_i} \tag{4.3}$$

kde |D| je celkový počet dokumentov, $tf_{i,j}$ je frekvencia výskytu termín *i* v dokumente *j* v dokumente a df_i je počet dokumentov, v ktorých sa vyskytol termín *i* [24]. Hodnota $tf_{i,j}$ sa počíta podľa vzorca:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4.4}$$

kde $n_{i,j}$ je počet výskytov termínu *i* v dokumente *j* [24]. Je to vlastne počet výskytov termínu v dokumente vydelený počtom všetkých termínov v tomto dokumente.

4.7 Vyhľadávanie kandidátnych termínov pre termíny a koncepty

Na vstupe sú zadané skúmané termíny, ku ktorým sa vyhľadávajú kandidátne termíny. Tie je možné vyhľadávať zo všetkých termínov v indexe alebo zo zadanej množiny termínov. Vyhľadávanie prebieha tak, že sa kontextový vektor každého skúmaného termínu porovná s kontextovými vektormi všetkých ostatných termínov. Porovnanie vektorov prebieha tak, že sa vypočíta skalárny súčin týchto vektorov a výsledná hodnota udáva podobnosť vektorov a teda aj termínov. Následne sa vyberú najlepšie kandidátne termíny. Je možné vybrať zadané množstvo najlepších termínov alebo termíny, ktoré sú podobnejšie ako zadaný prah. Na výstupe sú pre skúmané termíny vypísané kandidátne termíny, ktoré sú zoradené od najpodobnejšieho k najmenej podobnému. Voliteľne je možné vypísať aj hodnotu podobnosti každého kandidátneho termínu. Pokiaľ sa pre skúmaný termín nenašiel žiadny kandidátny termín (kvôli zadanému prahu), nie je tento skúmaný termín uvedený vo výsledkoch.

Vyhľadávanie kandidátnych termínov pre koncepty prebieha rovnako, ako vyhľadávanie kandidátnych termínov pre skúmané termíny, pretože koncepty sú reprezentované kontextovými vektormi rovnako ako termíny.

4.8 Vyhľadávanie kontextov pre skúmané termíny

Pre všetky zadané skúmané termíny sú kontexty vyhľadávané v SW1 korpuse. Vyhľadávajú sa v SW1 korpuse preto, lebo obsahuje články z Wikipédie, u ktorých je pravdepodobné, že sa termíny objavujú v zmysluplných vetách, a teda kontexty nesú užitočné informácie. Nevyberajú sa teda kontexty, ktoré nenesú žiadnu užitočnú informáciu. Takéto kontexty, ktoré nenesú žiadnu užitočnú informáciu, sa však nachádzajú vo Web 1T korpuse, ktorý bol ďalším potenciálnym zdrojom pre vyhľadávanie kontextov. Ten obsahuje aj také n-gramy, ktoré sú tvorené len termínom a jedným znakom, napríklad úvodzovkami, číslicou a podobne. Pracovať s takýmito kontextami je zbytočné a navyše to spomaľuje systém.

Korpus je postupne prechádzaný po vetách. Keď sa vo vete vyskytne nejaký hľadaný termín, vyberú sa z tejto vety všetky úseky, ktoré obsahujú tento termín a jeho kontext, ktorý má dĺžku jedno, dve, tri alebo štyri slová. V týchto úsekoch je potom skúmaný termín nahradený za žolíkový znak a takto vzniknutý reťazec je priradený ako kontext k danému termínu. Vybrané úseky vety môžu byť dlhšie ako päť slov, pretože aj viacslovný termín sa považuje za jedno slovo. Vo vzniknutom reťazci je potom aj viacslovný termín nahradený len jedným žolíkovým znakom, takže vzniknuté reťazce majú minimálnu dĺžku dve slová a maximálnu dĺžku päť slov. Toto je obmedzenie Web 1T korpusu,

ktorý obsahuje n-gramy, kde 1 <= n <= 5. Unigramy (1-gramy) však nemá zmysel skúmať, lebo by obsahovali iba termín a žiadny kontext. Nepridávajú sa také kontexty, ktoré oproti už vybraným kontextom obsahujú naviac len slová, ktoré neprešli filtrovaním, ktoré je popísané v kapitole 4.2.

4.9 Vyhľadávanie kontextov pre kandidátne termíny

Kontexty pre kandidátne termíny sa vyhľadávajú tak, že pred a za termín sa pridajú žolíkové znaky tak, aby vzniknutý reťazec mal dĺžku dva, tri, štyri alebo päť slov. Takýto kontext sa vytvorí pre každú pozíciu termínu v n-grame. Takto vzniknuté n-gramy sa vyhľadávajú vo Web 1T korpuse. Všetky nájdené n-gramy sa potom porovnajú s kontextami, ktoré sa našli v SW1 korpuse pre skúmané termíny. Odfiltrujú sa tie n-gramy, ktoré sa medzi kontextami skúmaných termínov nenachádzajú. Pre tie, ktoré sa tam nachádzajú, sa vypočíta ohodnotenie a následne sa priradia k príslušným kandidátnym termínom.

4.10 Ohodnotenie kandidátnych termínov

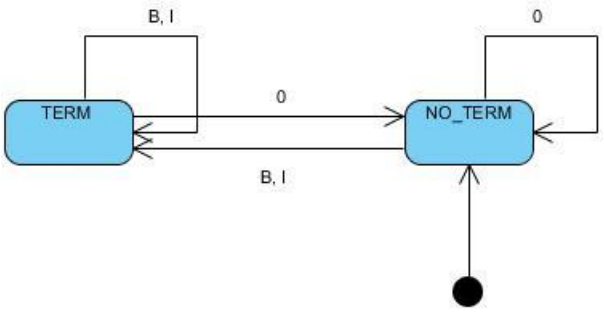
Ohodnotenie kandidátnych termínov prebieha pomocou porovnávania kontextov skúmaných termínov a kontextov ich kandidátnych termínov. Prejdú sa všetky kontexty a sčítajú sa ohodnotenia kontextov kandidátnych termínov, ktoré sa zároveň nachádzajú aj medzi kontextami skúmaného termínu. Kandidátne termíny sú následne usporiadané od najpodobnejšieho po najmenej podobný, pričom hodnotu podobnosti určuje súčet ohodnotení tých kontextov kandidátnych termínov, ktoré sú spoločné s daným skúmaným termínom.

5 Implementácia

V tejto kapitole je rozobraná implementácia jednotlivých modulov systému. Jednotlivé časti systému sú implementované v jazyku Java SE. Tieto programy sú pospájané pomocou skriptu v jazyku bash. Na vyhľadávanie vo Web 1T korpuse slúži nástroj Get1T, napísaný v jazyku C. Ako hlavný implementačný jazyk bola zvolená Java SE najmä preto, lebo v nej sú napísané knižnice Apache Lucene a semanticvectors, ktoré sa v systéme využívajú. Použité sú 64-bitové verzie Java Virtual Machine a prekladača gcc pre preklad jazyka C. To je spôsobené tým, že kvôli veľkosti vstupu vznikajú ako medzivýsledky súbory väčšie ako 4GB, ktoré 32-bitové verzie nie sú schopné spracovať. 64-bitové programy môžu tiež využiť viac operačnej pamäte, čo umožní spracovávať väčšie množstvo dát naraz a zrýchliť tak činnosť celého systému.

5.1 Extrakcia termínov

Termíny sú extrahované z SW1 korpusu. Celý korpus sa spracováva v jednom prechode. Každý nájdený termín sa zapisuje do výstupného súboru. Termíny sa hľadajú pre všetky zdroje v SW1 korpuse súčasne. Zdroje termínov v SW1 korpuse sú popísané v kapitole 3.2. Každý zdroj má vlastný stavový automat. Všetky tieto automaty sú zhodné. Stavové automaty sú dvojstavové. Prvý stav znamená, že sa spracováva termín a druhý, že sa nespracováva.



Obrázok 5.1 Stavový automat pre extrakciu termínov

Znaky B, I, 0 označujú termíny a ich význam je popísaný v kapitole 3.2. Pomocou týchto znakov je možné extrahovať aj viacslovné termíny.

V prípade, že je automat v stave TERM a príde znak B, aktuálny termín pre daný zdroj sa zapíše do výstupného súboru, zvýši sa frekvencia jeho výskytu o jedna a začne sa spracovávať ďalší termín. Okrem týchto zdrojov termínov sa budú za termíny považovať aj všetky podstatné mená, prídavné mená a slovesá. Nie všetky sa síce považujú za termíny, ale môžu pomôcť pri výpočte podobnosti termínov. Napríklad sloveso „drive“ sa často viaže s dopravnými prostriedkami. Takže ak sa toto sloveso bude vyskytovať pri dvoch rôznych termínoch, je možné usudzovať, že obidva termíny sú dopravné prostriedky a sú si teda významovo blízke. Slovesá, podstatné a prídavné mená

sa extrahujú na základe tagov, ktoré označujú ich slovný druh. Pokiaľ sa termín vyskytne vo viacerých zdrojoch súčasne, zapíše sa iba jeden výskyt termínu. Všetky písmená v termíne sú prevedené na malé. Keby neboli, mohol by sa termín vyskytnúť niekedy s veľkými písmenami a niekedy s malými. To by spôsobilo, že by bol tento termín spracovaný ako niekoľko rozdielnych termínov, čo by mohlo skresliť výsledky.

Termíny sa skladajú zo základných tvarov slov, ktoré sú v SW1 korpuse v treťom stĺpci. Základné tvary sú použité najmä preto, lebo v texte sa termín môže vyskytnúť vo viacerých formách (v angličtine napríklad množné číslo). Keby sa použili pôvodné tvary slov, tak by napríklad mohli byť nájdené termíny „car accident“ a „car accidents“, ktoré by boli považované za dva rozdielne termíny. Toto by spôsobovalo nepresnosti pri výpočte podobnosti termínov. Na druhú stranu niekedy môžu termíny obsahovať aj slová, ktoré nie sú v základnom tvare. Napríklad termín „Three Gorges Dam“ (priehrada Tri rokliny v Číne) obsahuje slovo „gorges“, ktoré je v množnom čísle. Pri použití základných tvarov slov sa potom zmení názov. Prípady, keď sa v termíne vyskytujú slová, ktoré nie sú v základnom tvare, sa objavujú najmä medzi vlastnými pomenovaniami (priehrady, mestá, ...). Neočakáva sa teda, že by sa často vyhľadávali. A keď sa budú vyhľadávať, dajú sa jednoduchým spôsobom previesť do požadovaného tvaru ešte pred vyhľadávaním. Preto a aj pre zjednodušenie vyhľadávania termínov a pre presnejší výpočet ich podobnosti sa použijú základné tvary, aby nedochádzalo k tomu, že jeden termín bude mať medzi nájdenými termínmi viac tvarov.

Pre každý termín sa zároveň počíta počet výskytov v texte. Pretože termínov sa vyskytuje veľké množstvo, po spracovaní každých 600 súborov SW1 korpusu príde k odstráneniu tých termínov, ktoré sa vyskytli menej ako päťkrát. Týmto sa odstráni veľa zbytočných termínov, čo šetrí pamäť a zrýchľuje vyhľadávanie, pretože zoznam termínov, ktorý je uložený a v ktorom je potrebné vyhľadávať, je menší. Tie termíny, ktoré prejdú týmto frekvenčným filtrom sú na konci spracovania SW1 korpusu filtrované filtrom, ktorý je popísaný v kapitole 4.2. Filtruje sa až na konci, lebo filtrovanie každého načítaného výskytu termínu by mohlo celý proces spomaliť (v korpuse sú desiatky až stovky miliónov výskytov termínov). Takto každý termín prejde filtrom maximálne jedenkrát.

Výstupom z extrakcie termínov sú dva súbory. Jeden obsahuje zoznam vyfiltrovaných termínov. Každý termín je na samostatnom riadku. Druhý súbor obsahuje termíny v jednotlivých článkoch. Každý termín je na samostatnom riadku. Zaznamenaný je každý výskyt termínu. To znamená, že pokiaľ sa termín v článku vyskytol desaťkrát, bude desaťkrát uvedený aj vo výstupnom súbore pri príslušnom článku. Zachováva sa aj poradie výskytov, pretože termíny sú zapisované v takom poradí, v akom sa našli v SW1 korpuse. Začiatok článku v súbore označuje tag <DOC názov_súboru>. V prípade, že názov článku nebolo možné získať, bude mať tag tvar <DOC>. Koniec súboru bude označený tagom </DOC>. Súbor však nebude mať formát súboru XML. Jeho veľkosť môže dosiahnuť niekoľko GB a parsovanie takto dlhého XML súboru by trvalo veľmi dlho. Navyše

súbor bude pri indexácii spracovaný jedným sekvenčným priechodom, takže použitie formátu XML nemá zmysel.

5.2 Tvorba indexu

Indexácia je implementovaná pomocou knižnice Apache Lucene. Na vstupe sú obidva súbory, ktoré vznikli pri extrakcii termínov. Indexer prechádza súbor, ktorý obsahuje termíny pre jednotlivé články a ukladá z neho do indexu len tie termíny, ktoré sa nachádzajú v druhom súbore. Na konci vznikne index v binárnom formáte používanom knižnicou Apache Lucene. Spracovanie vstupného súboru prebehne v jednom prechode. V indexe je každý článok v samostatnom dokumente. Dokument obsahuje pole s názvom článku a pole s termínmi z článku v rovnakom poradí a počte, v akom boli na vstupe. Vynechané sú len tie termíny, ktoré nie sú v zozname v druhom vstupnom súbore. V prípade, že na vstupe nie je uvedený názov článku, zostane pole s názvom prázdne.

5.3 Tvorba kontextových vektorov pre termíny

Kontextové vektory sa tvoria z indexu. Najprv sa pre každý termín v indexe vytvorí jedinečný náhodný vektor, ktorý bude pozostávať z niekoľkých hodnôt +1 a -1 a všetky ostatné hodnoty sú nulové. Počet +1 a -1, rovnako ako aj dimenziu vektoru, je možné nastaviť parametrami. Jedinečnosť vektorov sa overuje pomocou hašovacej funkcie. Takže nie sú povolené dva vektory s rovnakou hodnotu hašovacej funkcie, aj keď nie sú zhodné. Takéto porovnávanie je rýchlejšie ako porovnávanie celých vektorov. Navyše v pamäti nie je pri generovaní náhodných vektorov potrebné udržovať celé vektory, ale len hodnoty hašovacích funkcií, čo šetrí pamäť. Prvky vektorov budú usporiadané vzostupne predtým, než bude vypočítaná hašovacia funkcia. Keby neboli usporiadané, nebolo by možné porovnávať vektory pomocou hašovacej funkcie, pretože poradie prvkov vo vektore ovplyvňuje jej výsledok.

Následne sa začne počítanie kontextových vektorov. Prechádza sa celý index. Pre každý výskyt termínu sa zoberie jeho okolie, ktorého veľkosť je možné nastaviť parametrom a ku kontextovému vektoru toho termínu sa pripočítajú náhodné vektory termínov v tomto okolí. V tomto okolí sa ignorujú tie termíny, ktoré sú obsiahnuté v skúmanom termíne a tie, ktoré obsahujú skúmaný termín. Napríklad pre termíny „hercule“, „poirot“ a „hercule poirot“, sa pri termíne „hercule“ preskočí termín „hercule poirot“ a naopak pre termín „hercule poirot“ sa preskočia termíny „hercule“ aj „poirot“. Takto sa zabraňuje tomu, aby boli viacslovné termíny ovplyvnené tým, že sa vyskytujú spolu s termínmi, ktoré sú súčasťou tohto viacslovného termínu.

Keďže Wikipédia obsahuje veľké množstvo termínov, ich spracovanie v jednom prechode indexu by si vyžadovalo veľké množstvo pamäte na uchovanie všetkých kontextových vektorov. Preto sa kontextové vektory počítajú vo viacerých prechodoch. V jednom priechode sa spočítajú vždy

kontextové vektory pre určité množstvo termínov. Toto množstvo je možné nastaviť. Pri využití 4GB RAM sa v jednom priechode dá spracovať približne 200 000 termínov.

Náhodné vektory nie je potrebné udržiavať v pamäti všetky. Do pamäte sa načítajú len náhodné vektory z nasledujúcich tisíc dokumentov, ktoré sa majú práve spracovať. To umožňuje mať vysokú rýchlosť spracovania pri menších nárokoch na pamäť.

Výstupom z toho programu sú normalizované kontextové vektory, uložené v binárnom formáte. Rovnako v binárnom formáte zostanú uložené aj náhodné vektory pre všetky termíny a nenormalizované kontextové vektory. Binárny formát je zvolený preto, lebo na disku zaberá menej miesta ako textový formát a jeho spracovanie je rýchlejšie.

5.4 Tvorba kontextových vektorov pre koncepty

Za koncept je považovaný jeden článok z Wikipédie, ktorý je v indexe uložený ako jeden dokument. Koncept je reprezentovaný názvom príslušného článku vo Wikipédii. V názve konceptu sú, na rozdiel od termínov, ponechané veľké písmená, aby sa takto koncepty odlíšili od termínov. Kontextové vektory pre koncepty je možné tvoriť dvoma spôsobmi. Buď sa použijú normalizované kontextové vektory pre termíny, alebo náhodné vektory pre termíny. Výpočet prebieha v obidvoch prípadoch rovnako, rozdiel je teda iba v použitých vektoroch. Kontextové vektory pre koncepty sú vypočítané z vektorov (normalizovaných kontextových alebo náhodných) pre termíny, ktoré sa vyskytli v tomto koncepte a sú prítomné v indexe. Voliteľne je možné na vstup zadať zoznam termínov, ktoré sa majú brať do úvahy. Ak takýto zoznam nebude zadaný, použijú sa všetky termíny v indexe. Vektory sú pred pripočítaním ku kontextovému vektoru pre koncept váhované metódou tf-idf, ktorá je popísaná v kapitole 4.6 (vzorce 4.3 a 4.4). Po vypočítaní vektoru pre daný koncept bude tento vektor normalizovaný.

Počítanie kontextových vektorov pre koncepty prebieha v jednom priechode indexom. Toto je možné, pretože Apache Lucene index obsahuje počet výskytov termínov v jednotlivých dokumentoch, rovnako ako počet dokumentov, v ktorých sa daný termín vyskytol. To uľahčuje výpočet hodnoty tf-idf. Vektory pre jednotlivé termíny sa načítavajú podľa potreby dopredu, vždy pre určitý počet dokumentov. Dokumenty sa spracovávajú po jednom.

Výsledné kontextové vektory sú uložené v binárnom formáte s rovnakou štruktúrou, akú má súbor s kontextovými vektormi pre termíny. V tomto prípade sa neukladajú nenormalizované kontextové vektory, pretože sú zbytočné.

5.5 Vyhľadávanie kandidátnych termínov pre termíny a koncepty

Kandidátne termíny je možné vyhľadávať spomedzi zadaných termínov alebo medzi všetkými termínmi v indexe. Pri zadaní termínov existujú dve možnosti. Buď bude zadaný jeden zoznam termínov a v rámci toho zoznamu sa porovná každý termín s každým. Zoznam skúmaných termínov a potenciálnych kandidátnych termínov je teda rovnaký. Alebo je možné zadať dva zoznamy. Jeden obsahuje skúmané termíny a druhý potenciálne kandidátne termíny. Následne sa porovnávajú každý termín zo skúmaného zoznamu s každým termínom zo zoznamu potenciálnych kandidátov.

Termíny sa porovnávajú pomocou skalárneho súčinu ich kontextových vektorov. Keďže sa násobia normalizované vektory, výsledok bude rovnaký ako pri kosínovej vzdialenosti. Výsledok môže nadobudnúť hodnoty od -1,0 do 1,0, vrátane okrajových hodnôt. Výsledok 1,0 nastane v prípade, že sa vektor násobí so zhodným vektorom. Táto hodnota sa vyskytne, keď sa termín porovnáva sám so sebou. Hodnota -1,0 by vyšla v prípade, že by sa vektor násobil s vektorom, ktorého prvky by mali rovnakú veľkosť ale opačné znamienka ako pôvodný vektor.

Pre každý termín sa vytvorí zoznam kandidátnych termínov, ktorý je usporiadaný od najpodobnejšieho po najmenej podobný. Z toho zoznamu je možné vybrať prvých N kandidátov alebo kandidátov, ktorých podobnosť je väčšia ako zadaný prah. Tieto možnosti je možné aj skombinovať. V tomto zozname nie je uvedený skúmaný termín, pre ktorý je tento zoznam vytvorený, pretože by to nemalo zmysel, nakoľko by vždy prvým prvkom v zozname bol tento skúmaný termín.

Keďže skúmaných aj kandidátnych termínov môže byť potenciálne veľké množstvo, nie sú všetky skúmané termíny spracovávané súčasne. Vždy sa spracováva určité množstvo skúmaných termínov súčasne. Toto množstvo je možné nastaviť parametrom. Tiež je možné nastaviť parametrom to, aby sa pre každý skúmaný termín udržovalo maximálne toľko kandidátov, koľko najviac sa má vypísať. Toto pomáha šetriť pamäť, ale môže to spomaliť vyhľadávanie kandidátov. Pamäť to šetrí tým, že sa v nej neudržujú kandidátne termíny, ktoré sa aj tak nakoniec nevypíšu na výstup. Spomalenie vyhľadávania bude spôsobené dodatočnou réžiou, ktorá je spojená s vyhľadaním a odstránením posledného prvku v zozname kandidátov. V prípade, že je zadaná minimálna hodnota podobnosti, nebudú sa kandidáti s menšou hodnotu pravdepodobnosti do zoznamu ukladať vôbec, čím sa opäť ušetrí pamäť. Tento krát nedôjde k spomaleniu systému, pretože takýto kandidát nie je vôbec do zoznamu pridaný a nie je ho teda potrebné odstraňovať.

Vyhľadávanie kandidátov pre koncepty prebieha rovnako ako pri vyhľadávaní kandidátov pre termíny. Môžu byť zadané dva zoznamy. Jeden obsahuje skúmané koncepty a druhý potenciálne kandidátne termíny. V prípade, že nie je niektorý z nich zadaný, prípadne nie je zadaný ani jeden, berú do úvahy všetky koncepty, respektíve termíny, v indexe.

Výstupom z toho modulu sú kandidátne termíny pre jednotlivé skúmané termíny, respektíve koncepty. Uložené sú v jednoduchom textovom súbore, ktorý bude mať na začiatku každého riadku skúmaný termín alebo koncept. Za ním sú tabulátormi oddelení jednotliví kandidáti. Je možné vypísať aj číselnú hodnotu podobnosti pre jednotlivých kandidátov. V takom prípade je táto hodnota vypísané za každým kandidátom a oddelená od neho tabulátorom.

5.6 Vyhľadávanie kontextov pre skúmané termíny

Kontexty sa vyhľadávajú v SW1 korpuse. V kontextoch sa, na rozdiel od extrakcie termínov, používajú pôvodné tvary slov, ktoré sú uvedené v prvom stĺpci. Je to z toho dôvodu, že Web 1T korpus obsahuje slová v pôvodnom tvare, v akom sa vyskytli na webových stránkach.

Spracovanie prebieha po vetách. Vety, ktoré neobsahujú žiadny zaindexovaný termín, sú preskočené. Vyberú sa štyri slová pred termínom a štyri slová za ním, pokiaľ ich je toľko vo vete. Inak sa zoberú len slová po koniec, respektíve od začiatku, vety. Z týchto slov sú potom tvorené kontexty. Celková dĺžka kontextu je jedno, dve, tri alebo štyri slová plus sa pridá jeden žolíkový znak pre termín. Poradie slov zostane rovnaké ako v pôvodnej vete. Výskyt termínu v kontexte sa označí žolíkovým znakom "<*>". Takto upravené kontexty potom umožňujú ich jednoduché porovnávanie, keď stačí porovnávať zhodu reťazcov, pretože ten istý kontext pre dva rozdielne termíny vyzerá rovnako. Každý nový kontext musí pridať nejakú užitočnú informáciu. Nevyberajú sa preto kontexty, ktoré oproti už existujúcim pridávajú len neužitočné slová, ako napríklad určité a neurčité členy, spojky a podobne. Za neužitočné slová sa budú považovať tie, ktoré neprejdú filtrom popísaným v kapitole 4.2. Tieto slová sa však neignorujú úplne. V prípade, že sa ku kontextu majú pridať dve alebo viac slov a aspoň prvé (ak sa pridávajú na začiatok) alebo posledné (ak sa pridávajú na koniec) z nich prešlo filtrom, pridajú sa všetky tieto slová.

Všetky takto vzniknuté kontexty sú priradené k príslušným termínom a uložené do súborov. Kontexty sa ukladajú do určeného adresáru, ktorý obsahuje podadresáre. Každý podadresár je pomenovaný prvými dvoma znakmi termínov, ktoré obsahuje. Toto je možné, pretože všetky termíny v indexe sú aspoň tri znaky dlhé. V každom tomto podadresári sú súbory. Názvy súborov sú rovnaké ako termíny. Každý súbor obsahuje kontexty, ktoré boli nájdené pre daný termín.

Okrem kontextov pre každý termín sa ukladajú aj všetky nájdené pravé a ľavé kontexty. Ľavý kontext je časť kontextu, ktorá sa nachádza pred termínom a pravý kontext je časť, ktorá sa nachádza za termínom. Pravé a ľavé kontexty sú uložené do súborov podľa ich dĺžky. Ukladajú sa len pravé a ľavé kontexty, bez žolíkových znakov namiesto termínu. Pravé a ľavé kontexty budú slúžiť na filtrovanie kontextov pre kandidátne termíny.

5.7 Vyhľadávanie kontextov pre kandidátne termíny

Kontexty pre kandidátne termíny sa vyhľadávajú pomocou nástroja Get1T[5] vo Web 1T korpuse. Jedná sa o program, ktorý je určený na predspracovanie a dopytovanie Web 1T korpusu. Je šírený pod GPL licenciou. Get1T očakáva na vstupe súbor, ktorý obsahuje n-gramy zhodnej dĺžky. Každý n-gram je na samostatnom riadku. Parametrom sa zadáva dĺžka n-gramov. Môže byť v rozmedzí od dva do päť. V n-gramoch môžu byť slová alebo žolíkové znaky "<*>". Každý takýto znak zastupuje jedno ľubovoľné slovo. Výstupom z Get1T je súbor, v ktorom budú všetky nájdené n-gramy spolu s počtom ich výskytov. V tomto súbore sú žolíkové znaky nahradené za skutočné slová. Na výstupe sa tiež nachádza súbor, ktorý obsahuje celkový počet n-gramov vo Web 1T korpuse pre danú dĺžku n. Get1T vyhľadáva n-gramy v jednom prechode Web 1T korpusom, preto je potrebné načítať všetky zadané n-gramy do operačnej pamäte. Get1T nevie vyhľadávať v unigramoch, pretože súbor, ktorý obsahuje unigramy je v nekomprimovanom formáte, s ktorým Get1T nedokáže pracovať. To však nevadí, pretože tento formát je rovnaký, ako výstupný formát z Get1T, takže nepotrebuje žiadne predspracovanie a je možné vyhľadávať unigramy priamo v tomto súbore.

Na vstup Get1T sa zadá súbor, ktorý obsahuje n-gramy. Tieto n-gramy vždy pozostávajú z kandidátneho termínu doplneného žolíkovými znakmi tak, aby mal daný n-gram správnu dĺžku. Pre každý termín sa vytvorí toľko n-gramov, aby bol termín na každej možnej pozícii v n-grame. Toto sa zopakuje pre všetky kandidátne termíny. Vytvoria sa štyri súbory pre 2-gramy, 3-gramy, 4-gramy a 5-gramy. Pre každý z týchto súborov sa spustí program Get1T a výsledkom sú štyri súbory, pričom každý obsahuje kontexty kandidátnych termínov rôznej dĺžky spolu s frekvenciou ich výskytu.

Tieto kontexty bude treba priradiť k správnym kandidátnym termínom. Je potrebné ich aj vyfiltrovať, pretože veľké množstvo kontextov by spomaľovalo výpočet podobnosti. Ponechajú sa iba tie kontexty pre skúmané termíny, ktoré sa našli v SW1 korpuse. K tomuto filtrovaniu poslúžia pravé a ľavé kontexty, ktoré boli nájdené v predchádzajúcom module. Tie sa budú tiež vyhľadávať vo Web 1T korpuse pomocou programu Get1T. Avšak súbory s týmito kontextami sú príliš veľké a Get1T by na ich spracovanie potreboval veľa pamäte, ktorá nemusí byť dostupná. Preto sú tieto kontexty rozdelené do viacerých súborov, ktoré sa spracujú samostatne. Následne sú výsledky pospájané tak, aby vznikli tri súbory, kde každý obsahuje kontexty dĺžky dva, tri a štyri. Kontexty dĺžky jedna budú vyhľadané samostatne, priamo vo Web 1T korpuse, kde sú uložené v rovnakom formáte, v akom je výstup z programu Get1T. Kontexty nemôžu byť dlhšie ako štyri slová, pretože maximálna dĺžka n-gramov je päť slov a okrem kontextu je potrebné mať v n-grame aj termín.

Keďže v nájdených n-gramoch nie je informácia o tom, pre ktorý termín je daný n-gram, je potrebné ho najprv priradiť ku správnemu kandidátnemu termínu. Zoberú sa všetky podfrázy

[5] http://get1t.sourceforge.net/

z nájdeného n-gramu s dĺžkou 1 až n-1, kde n predstavuje počet slov v n-grame, pretože termín môže byť v n-grame na ľubovoľnej pozícii a môže ho tvoriť jedno až n-1 slov (n je počet slov v n-grame). Všetky tieto podfrázy sa vyhľadajú v zozname kandidátnych termínov. Ak sa podfráza nájde v zozname kandidátov, n-gram sa rozdelí na tri časti, a to termín, ľavý a pravý kontext. Je zaznamenaný kandidátny termín, ľavý a pravý kontext, dĺžka ľavého a pravého kontextu, dĺžka n-gramu a frekvencia výskytu n-gramu. Z jedného n-gramu môžu vzniknúť kontexty k viacerým termínom, pokiaľ n-gram obsahuje viac termínov.

Keď sú kontexty takto spracované pristúpi sa k druhej fázy, v ktorej sú filtrované a je počítané ohodnotenie kontextov. Filtrácia prebieha tak, že sa jednotlivé ľavé a pravé kontexty kandidátnych termínov sú vyhľadávané medzi ľavými a pravými kontextami skúmaných termínov. Pravé a ľavé kontexty tvoria jeden zoznam, takže sa nerozlišuje, či sa jedná o pravý alebo ľavý kontext. Tie kontexty, pri ktorých nie je nájdený pravý alebo ľavý kontext, sa odstránia. Pre tie kontexty, ktoré zostanú je vypočítané ohodnotenie, ktoré vznikne podielom hodnoty pointwise mutual information (PMI) pre n-gram a hodnoty self-information (SI) pre pravý a ľavý kontext. Vzorce, podľa ktorých sa vypočítajú hodnoty PMI a SI sú vzorce 2.7, respektíve 2.8. Pravdepodobnosť výskytu sa vypočíta ako podiel frekvencie výskytu daného n-gramu podelený celkovým počtom n-gramov, pričom n označuje dĺžku n-gramu. Self-information sa vypočíta ako súčet self-information pravého a ľavého kontextu. Súčet je zvolený preto, lebo hodnota SI je počítaná pomocou logaritmu.

Na výstupe je pre každý kandidátny termín uvedený zoznam kontextov, ktoré sa našli vo Web 1T korpuse a prešli popísaným filtrom. Pri každom kontexte bude uvedené aj jeho ohodnotenie. Formát výstupu bude rovnaký, ako formát popísaný v kapitole 5.6.

5.8 Ohodnotenie kandidátnych termínov

Ohodnotenie kandidátnych termínov prebieha pomocou porovnávania kontextov. Na vstupe sa nachádzajú kontexty pre skúmané termíny a kontexty pre kandidátne termíny. Na vstupe je tiež zoznam, ktorý obsahuje kandidátne termíny priradené k jednotlivým skúmaným termínom.

Postupne sa prechádzajú skúmané termíny. Neprechádzajú sa po jednom, ale spracuje sa ich viac zároveň. Tento počet je možné zadať parametrom. Vytvorí sa zoznam kandidátnych termínov všetkých týchto spracovávaných skúmaných termínov. Tento zoznam sa prechádza po jednom termíne a každý kandidátny termín sa porovnáva s tými spracovávanými skúmanými termínmi, ku ktorým patrí. Takéto spracovávanie po skupinách je kompromisom medzi spotrebou pamäte a rýchlosťou spracovávania. Uchovávanie kontextov pre všetky skúmané termíny by bolo veľmi pamäťovo náročné a ich spracovávanie po jednom by bolo zas príliš pomalé. Keď sa spracováva viac skúmaných termínov súčasne, zmenšuje sa počet načítavaných termínov, pretože niektoré kandidátne termíny sa môžu vyskytnúť pri viacerých skúmaných termínoch. Takto stačí načítať tento kandidátny

termín iba raz (prípadne viackrát, ak sa nachádza pri skúmaných termínoch, ktoré sa spracovávajú v rozdielnych prechodoch) .

Porovnávanie kontextov prebieha tak, že sa porovnáva zhodnosť reťazcov. Toto umožňuje štruktúra kontextov, v ktorých je výskyt termínu nahradený žolíkovým znakom. Rovnaké kontexty pre rôzne termíny teda vyzerajú rovnako. Ak sa nejaký kontext kandidátneho termínu nachádza aj medzi kontextami skúmaného termínu, pripočíta sa ohodnotenie tohto kontextu pre kandidátny termín k ohodnotenie kandidátneho termínu. Kandidátne termíny pre každý skúmaný termín sú potom usporiadané podľa ich ohodnotenia od najvyššej hodnoty po najmenšiu. Výsledky sú vypísané do súboru, kde je na každom riadku jeden skúmaný termín a za ním kandidátne termíny, prípadne aj s ohodnotením. Jednotlivé termíny, prípadne ohodnotenia, sú od seba oddelené tabulátormi.

6 Metódy vyhodnotenia výsledkov

Výsledné blízke termíny získané metódou lexikálnej substitúcie aj medzivýsledky získané metódou Random Indexing sa porovnajú s niekoľkými existujúcimi dátovými sadami. Všetky tieto sady boli vytvorené ručne, takže reprezentujú ľudské chápanie podobnosti termínov, ku ktorému sa systém snaží priblížiť.

Na vyhodnotenie podobnosti sa použijú štandardné metriky. Prvou metrikou bude korelácia. Korelácia sa počíta ako Pearsonov korelačný koeficient podľa vzorca:

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2)E^2(Y)}} \tag{6.1}$$

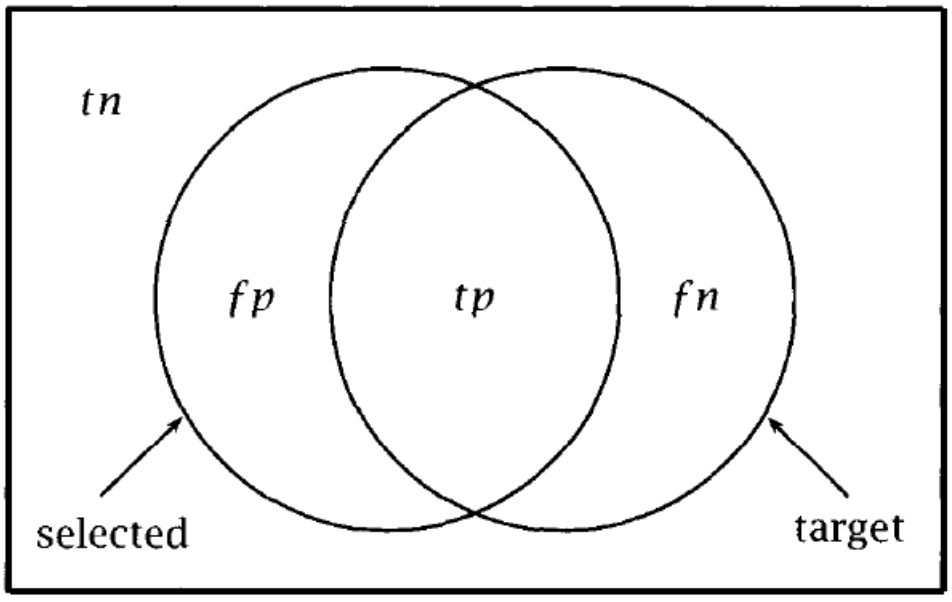
kde X sú hodnoty z prvého súboru, Y sú hodnoty z druhého súboru, E(N) je priemerná hodnota súboru N, $E^2(N)$ je druhá mocnina priemernej hodnoty súboru N, $E(N^2)$ je priemerná hodnota z druhých mocnín prvkov zo sady N [25].

Ďalšou metrikou je presnosť a pokrytie. Presnosť a pokrytie sa počítajú podľa vzorcov:

$$\text{presnosť} = \frac{tp}{tp + fp} \tag{6.2}$$

$$\text{pokrytie} = \frac{tp}{tp + fn} \tag{6.3}$$

kde *tp* znamená počet termínov, ktoré systém správne označil ako podobné, *fp* znamená počet termínov, ktoré systém nesprávne označil ako podobné a *fn* značí počet termínov, ktoré systém chybne neoznačil ako podobné [19].



Obrázok 6.1 Význam hodnôt tp, fp, fn pri výpočte presnosti a pokrytia

Obrázok 6.1 zobrazuje význam hodnôt *tp, tf, fp*. Prevzatý je z [19]. Množina *selected* označuje termíny, ktoré vybral systém ako podobné a množina *target* označuje termíny, ktoré boli vybrané ako podobné v porovnávaných sadách.

Presnosť počíta pomer správne označených podobných termínov ku všetkým termínom, ktoré boli označené ako podobné. Pokrytie počíta pomer správne označených podobných termínov ku všetkým podobným termínom z porovnávanej sady.

Pri výpočte sa použije metóda micro-averaging. To znamená, že sa hodnoty *fp, tp* a *fn* vypočítajú globálne z celého súboru [19]. Použité budú vzorce:

$$\text{presnosť} = \frac{\sum_{i=1}^{m} |tp_i|}{\sum_{i=1}^{m} (|tp_i| + |fp_i|)} \tag{6.4}$$

$$\text{pokrytie} = \frac{\sum_{i=1}^{m} |tp_i|}{\sum_{i=1}^{m} (|tp_i| + |fn_i|)} \tag{6.5}$$

kde *m* je počet skúmaných termínov, pre ktoré systém hľadal podobné termíny, tp_i , fp_i a fn_i sú hodnoty *tp, fp* a *fn* pre i-tý skúmaný termín [1].

6.1 WordNet

WordNet je rozsiahla lexikálna databáza anglického jazyka, ktorá bola vytvorená pod vedením Georga A. Millera. Termíny sú uchovávané v skupinách synoným, ktoré vyjadrujú určité koncepty. Tieto skupiny sa volajú synsety. WordNet napríklad obsahuje synset, ktorý obsahuje slová „world“, „Earth“, „globe“, a ďalší synset, ktorý obsahuje slová „world“, „universe“, „existence“, „cosmos“ a „macrocosm“. Obidva sysnety obsahujú slovo „world“ a jeho synonymá, ale každý synset predstavuje iný koncept.

Synsety sú poprepájané na základe konceptovo-sémantických a lexikálnych vzťahov. WordNet rozlišuje medzi podstatnými a prídavnými menami, slovesami a príslovkami, pretože sa správajú podľa rozdielnych gramatických pravidiel. Pokiaľ má termín viac významov, je prítomný v každom synsete, ktorý reprezentuje niektorý význam tohto termínu [7]. Synsety pre podstatné mená sú usporiadané v hierarchickej stromovej štruktúre, ktorá je založená na vzťahoch hyperonymum-hyponymum. Jedná sa o vzťahy medzi nadradenými a podradenými slovami. Napríklad slovo „jablko“ je hyponymom slova „ovocie“ (každé jablko je ovocie, ale nie každé ovocie je jablko). Naopak slovo „ovocie“ je hyperonymom slova „jabko“. Okrem tohto základného usporiadania sú v štruktúre aj ďalšie odkazy medzi synsetmi, ktoré vyjadrujú ďalšie vzťahy medzi konceptmi, ako napríklad antonymá.

Spočíta sa presnosť a pokrytie výsledkov voči WordNetu. Z WordNetu sa vyberú pre každý skúmaný termín ako synonymá tie slová, ktoré sa spolu s daným skúmaným termínom vyskytli aspoň v jednom synsete. Presnosť a pokrytia sa počítajú podľa vzorcov 6.4 a 6.5.

6.2 Asociačný test

Asociačný test je test, pri ktorom je človeku povedané slovo a on má odpovedať slovom, ktoré ho prvé napadne. Na testovanie systému sa využíva Edinburgh Associative Thesaurus (EAT)[6]. Obsahuje približne 8400 podnetov [37]. Podnet je slovo, pre ktoré mal testovací subjekt nájsť najbližšie slovo. Každý podnet bol daný 100 rôznym ľuďom. Každý človek, ktorý sa zúčastnil testu dostal 100 rôznych podnetov. Na každý podnet mal napísať prvé slovo, ktoré ho napadlo. Väčšina ľudí dokončila úlohu za 5 až 10 minút. Odpovede osôb, ktoré nevyplnili viac ako 25% odpovedí, neboli brané do úvahy. Zo zvyšných bol vytvorený EAT [37]. EAT obsahuje zoznam podnetov a ku každému podnetu uvádza zoznam slov, ktoré testovacie subjekty uviedli ako odpovede na daný podnet.

Hodnotí sa presnosť a pokrytie výsledkov systému voči asociačnému testu. Na výpočet presnosti a pokrytia sú použité vzorce 6.4 a 6.5.

6.3 WordSimilarity-353 Test Collection

WordSimilarity-353 Test Collection[7] obsahuje dve množiny párov anglických slov spolu s hodnotou ich podobnosti, ktorú im priradili ľudia [39]. Táto sada sa často používa trénovanie a testovanie algoritmov na výpočet sémantickej blízkosti.

Prvá množina obsahuje 153 párov, ktoré hodnotilo 13 ľudí. Druhá množina obsahuje 200 párov, ktoré hodnotilo 16 ľudí. Celkovo teda kolekcia obsahuje 353 párov, z čoho vznikol aj názov. Testovacie osoby mali určiť podobnosť slov na stupnici od 0 po 10, kde 0 znamená úplne rozdielne slová a 10 znamená veľmi podobné alebo identické slová. Hodnota nemusela byť celé číslo. V prípade, že boli zadané slová antonymá, ľudia dostali pokyn, aby tieto slová brali skôr ako podobné. To znamená, že antonymá majú v tejto sade vyššie hodnoty podobnosti. Pre každý pár slov sa následne určila priemerná hodnota podobnosti [39]. Aj napriek subjektívnej povahe takto získaných dát, dosahovala korelácia medzi hodnotami jednotlivých testovacích subjektov a priemerom 0.79 [20].

Počíta sa korelácia výsledkov systému s touto dátovou sadou podľa vzorca 6.1. Prehľad korelácie výsledkov niektorých algoritmov pre výpočet sémantickej blízkosti termínov s WordSimilarity-353 Test Collection prevzatý z [20] je uvedený v tabuľke 6.1.

[6] http://www.eat.rl.ac.uk/
[7] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

Algoritmus	Korelácia s WordSimilarity-353
WordNet	0,33 – 0,35
Roget's Thezaurus	0,55
LSA	0.56
WikiRelate!	0,19 – 0,48
ESA	0,75

Tabuľka 6.1 Korelácia výsledkov niektorých algoritmov s WordSimilarity-353

6.4 Metódy vyhodnotenia podobnosti konceptov

Podobnosť konceptov s termínmi sa vyhodnotí na päťdesiatich konceptoch, ktoré sú uvedené v [9]. Tieto koncepty sú uvedené v tabuľke 6.2.

Sex	Pornography	Love	Politics	Book
Earth	Map	Computer	Cat	Game
Dictionary	Ejaculation	Lesbian	Sport	Cancer
God	Virus	Animal	Heroin	Horse
Film	Human	Condom	Snake	Flower
Evolution	Beer	Statistics	Religion	Communication
Management	Insomnia	Erection	Death	Earthquake
People	Insurance	Graffiti	Research	Puberty
Cannabis	Marriage	Gun	Socialism	Narcissism
Twilight	Sleep	Architecture	Stroke	Leaf

Tabuľka 6.2 Koncepty, na ktorých sa bude vyhodnocovať systém

Podobné termíny pre tieto koncepty sa porovnajú so synonymami, ktoré budú získané z voľne dostupného on-line tezauru[8], v ktorom sa tieto koncepty namapujú na správne významy slov. Zo synoným sa pre zjednodušenie vyberú len podstatné mená.

Koncepty boli vyberané tak, aby sa jednalo o jednoslovné termíny. Zároveň pre tieto koncepty musí existovať článok na Wikipédii, aby v indexe existoval pre daný termín koncept.

Presnosť a pokrytie sa počítajú podľa vzorcov 6.4 a 6.5.

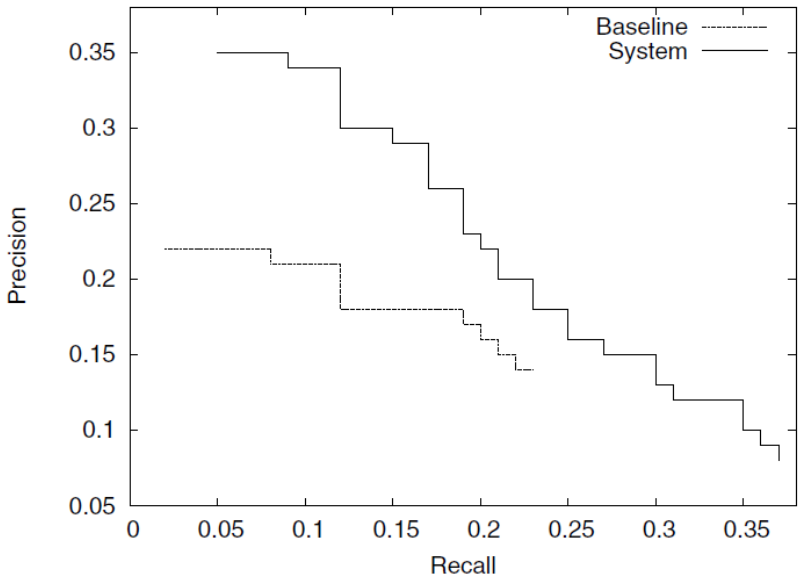
[8] http://thesaurus.com/

7 Vyhodnotenie výsledkov

V tejto kapitole sú uvedené výsledky jednotlivých experimentov spolu s ich zhodnotením. Rovnako je uvedená aj časová náročnosť jednotlivých častí systému. Výsledky, keď sa brali do úvahy všetky nájdené podobné termíny, pre Random Indexing a lexikálnu substitúciu sa môžu mierne líšiť, pretože pri lexikálnej substitúcii sa ignorujú tie termíny, pre ktoré sa nenašli žiadne kontexty a nedá sa pre ne určiť podobnosť metódou lexikálnej substitúcie. Tieto prípady sa však týkajú menej ako 1% skúmaných termínov, takže výsledky nie sú týmto veľmi ovplyvnené.

Pri experimentoch, keď sa počítala presnosť a pokrytie sa získalo niekoľko týchto hodnôt. Tieto hodnoty sa získali tak, že sa bral do úvahy iba prvý najpodobnejší termín, ktorý vybral systém, potom prvé dva najpodobnejšie termíny, a tak ďalej až prvých desať najpodobnejších termínov. Pre všetky tieto hodnoty sa brali do úvahy všetky podobné termíny z porovnávanej sady. Zo získaných hodnôt presnosti a pokrytia sa zostavila tabuľka a graf závislosti presnosti od pokrytia. Grafy závislosti presnosti od pokrytia budú uvedené v rovnakom tvare, v akom sú zobrazené v [9].

Zistené hodnoty presnosti a pokrytia budú porovnávané s hodnotami z [9]. Tieto hodnoty sú uvedené v grafe 7.1. V [9] boli ako skúmané termíny zvolené koncepty z tabuľky 6.2. Všetky tieto termíny predstavovali jeden článok vo Wikipédii. Ako kandidátne termíny k týmto konceptom boli zvolené tie termíny, ktoré boli obsiahnuté v hypertextových odkazoch, ktoré ukazovali na daný článok. Všetky takto získané kandidátne termíny boli zoradené podľa frekvencie ich výskytov v odkazoch na daný článok. Z takto zoradených termínov sa vypočítalo pokrytie a presnosť, ktoré sú v grafe 7.1 uvedené ako krivka *Baseline*. Následne boli kandidátne termíny preusporiadané metódou lexikálnej analýzy a pre výsledky bolo opäť spočítaných niekoľko hodnôt pokrytia a presnosti, ktoré sú v grafe 7.1 uvedené ako krivka *System*.



Obrázok 7.1 Závislosť presnosti od pokrytia prevzatý z [9]

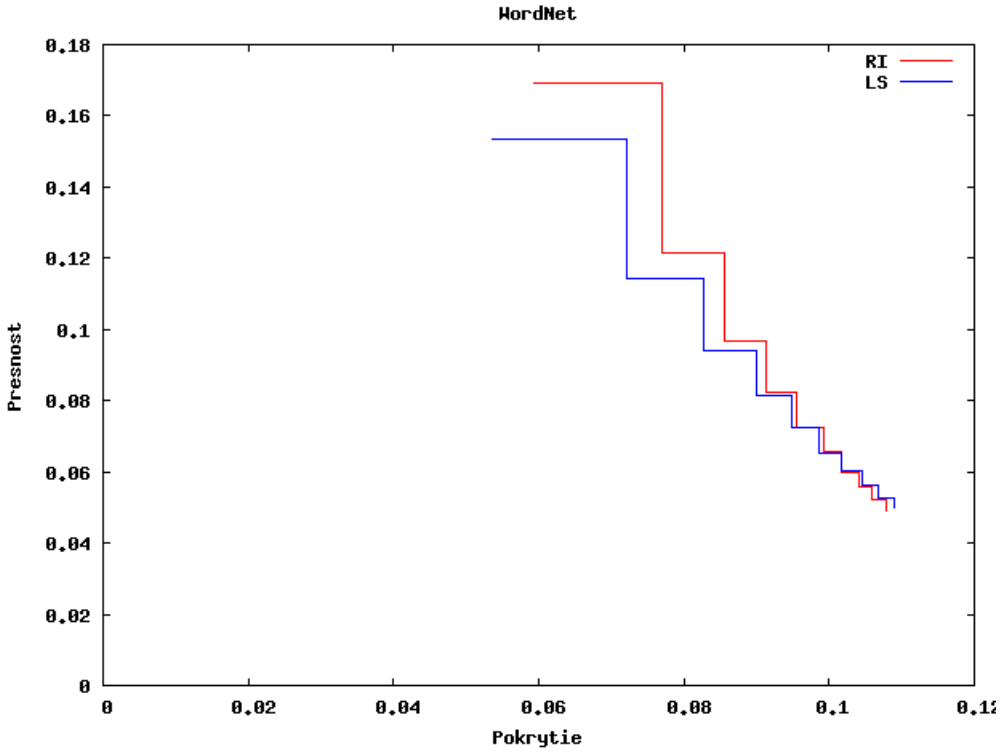
V grafoch v tejto kapitole je skratkou *RI* označený Random Indexing, *LS* znamená lexikálna substitúcia a *cw* je veľkosť kontextového okna.

7.1 WordNet

Celkovo sa vybralo 92 072 termínov, ktoré sa nachádzajú vo WordNete a boli zaindexované systémom. Následne sa tieto termíny porovnávali každý s každým. Systém vybral ako kandidátne termíny tie, ktoré mali so skúmaným termínom pri metóde Random Indexing podobnosť aspoň 0,5. Následne sa skúmala hodnota presnosti a pokrytia voči synonymám, ktoré sú prítomné vo WordNete. Vhodné by bolo okrem synoným skúmať aj niekoľko ďalších najpodobnejších termínov. K tomu by bolo potrebné spočítať podobnosť každého termínu s každým, pretože inak sa okrem synoným ďalšie najpodobnejšie termíny vyhľadať nedajú. Tento výpočet by však bolo veľmi časovo náročný. Preto sa spočíta presnosť a pokrytie iba voči synonymám. Tabuľka 7.1 obsahuje hodnoty presnosti a pokrytia, keď sa bral do úvahy iba jeden najpodobnejší termín, až pokiaľ sa bralo do úvahy 10 najpodobnejších termínov. Na konci sú uvedené hodnoty pre všetky nájdené podobné termíny. WordNet obsahuje pre termíny menej synoným, obvykle asi štyri až päť synoným, pre niekoľko málo termínov však obsahuje aj viac ako 10 synoným.

Podobné termíny	Random Indexing		Lexikálna substitúcia	
	pokrytie	presnosť	pokrytie	presnosť
1	0,059278	0,169382	0,053544	0,153521
2	0,076918	0,121275	0,072123	0,114116
3	0,085530	0,096757	0,082747	0,093959
4	0,091231	0,082179	0,089948	0,081327
5	0,09562é	0,072507	0,094956	0,072283
6	0,099187	0,065521	0,098712	0,065466
7	0,101748	0,059952	0,101681	0,060149
8	0,104057	0,055597	0,104450	0,056034
9	0,105972	0,051989	0,106810	0,052618
10	0,107799	0,049044	0,108893	0,049751
Všetky	0,141284	0,015023	0,140368	0,015072

Tabuľka 7.1 Hodnoty pokrytia a presnosti pri zadanom prahu 0,5



Obrázok 7.2 Graf závislosti presnosti od pokrytia výsledkov systému voči WordNetu

Z grafu 7.2 je vidieť, že výsledky systému dosahujú približne polovičné až tretinové hodnoty presnosti a pokrytia oproti výsledkom uvedeným v [9], ktoré sú uvedené na grafe 7.1. Hodnoty pre Random Indexing sú mierne lepšie, ale tento rozdiel sa postupne stráca, keď sa berie do úvahy viac najpodobnejších termínov.

Hodnoty presnosti a pokrytia by mohli byť lepšie, keby sa okrem synoným brali do úvahy aj iné podobné termíny, ktoré patria do rovnakej domény. Systém má totiž tendenciu vyberať ako najpodobnejšie termíny tie, ktoré nie sú synonymá, ale patria do rovnakej domény. Napríklad pre termín „harpoon“[9] sa našli ako najpodobnejšie termíny „exocet“, „tomahawk“, respektíve „stinger“. Takže pre pomenovanie riadenej strely sa našli ako najpodobnejšie termíny pomenovania iných riadených streiel. Tieto termíny sú si blízke, ale nejedná sa o synonymá. Pre viac špecifické termíny (názvy miest, ...) sú výsledky väčšinou lepšie ako pre všeobecnejšie termíny (napríklad „man“, „wind“, ...). Pre termín „wind“ sa našli ako podobné termíny „drop“, „turn“, respektíve „winds“, „light“, ktoré sa nedajú považovať za veľmi podobné (okrem množného čísla termínu „wind“). To je spôsobené najmä tým, že všeobecnejšie termíny sa vyskytujú v kontextoch, ktoré sú často všeobecné alebo sa vyskytujú vo väčšom množstve kontextov, ktoré sa týkajú rôznych domén. Špecifickejšie termíny sa obvykle vyskytujú v kontextoch, ktoré sa viac týkajú daného termínu.

[9] tu vo význame riadená strela

7.2 Asociačný test

Pri asociačnom teste sa počítala presnosť a pokrytie výsledkov voči zadanému asociačnému testu. Asociačný test EAT obsahuje 8211 skúmaných termínov, ku ktorým sú vybrané asociácie. V teste je celkovo 22 764 termínov, ktoré sú priradené ako asociácie k aspoň jednému skúmanému termínu. Experiment s asociačným testom prebiehal tak, že pre každý skúmaný termín z asociačného testu sa vyberú podobné termíny spomedzi všetkých termínov, ktoré boli priradené ako asociácie k nejakému termínu. Pre každý termín sa teda vyberajú podobné termíny spomedzi 22 764 termínov.

Nasledujúce tabuľky ukazujú hodnoty pokrytia a presnosti pre prvý až prvých desať najpodobnejších termínov. V tomto experimente sa ako kandidátne termíny brali do úvahy všetky termíny, ktorých podobnosť so skúmaným termínom pri metóde Random Indexing bola väčšia ako hodnota 0,5. Pri metóde Random Indexing bola veľkosť kontextového okna 4, 6 a 10 termínov. To znamená, že pre každý termín sa pri počítaní kontextových vektorov brali do úvahy 4, 6 alebo 10 najbližších termínov. V poslednom riadku tabuľky sú uvedené hodnoty presnosti a pokrytia pre všetky nájdené podobné termíny.

Pri tomto experimente sa počítalo iba s tými skúmanými termínmi, pre ktoré sa našiel aspoň jeden kandidátny termín, ktorého podobnosť s daným skúmaným termínom bola viac ako 0,5. Pre väčšie veľkosti kontextového okna bolo vybraných viac skúmaných termínov (4 339, resp. 4 906, resp. 5696), čo je pravdepodobne spôsobené tým, že zväčšenie uvažovaného kontextu termínu zvyšuje pravdepodobnosť, že v tomto kontexte sa nachádzajú rovnaké termíny ako pri iných termínoch. Toto následne spôsobuje vyššie hodnoty podobnosti medzi termínmi.

Podobné termíny	**Random Indexing**		**Lexikálna substitúcia**	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**
1	0,004317	0,173437	0,005487	0,220327
2	0,006606	0,146032	0,008685	0,191932
3	0,008316	0,132249	0,010826	0,172096
4	0,009736	0,124077	0,012525	0,159561
5	0,010901	0,11773	0,013719	0,148144
6	0,011884	0,112515	0,014672	0,138913
7	0,012679	0,107790	0,015561	0,132308
8	0,013417	0,104232	0,016307	0,126712
9	0,014093	0,101294	0,016927	0,121694
10	0,01469	0,098650	0,017524	0,117713
Všetky	0,03131	0,035410	0,031312	0,035358

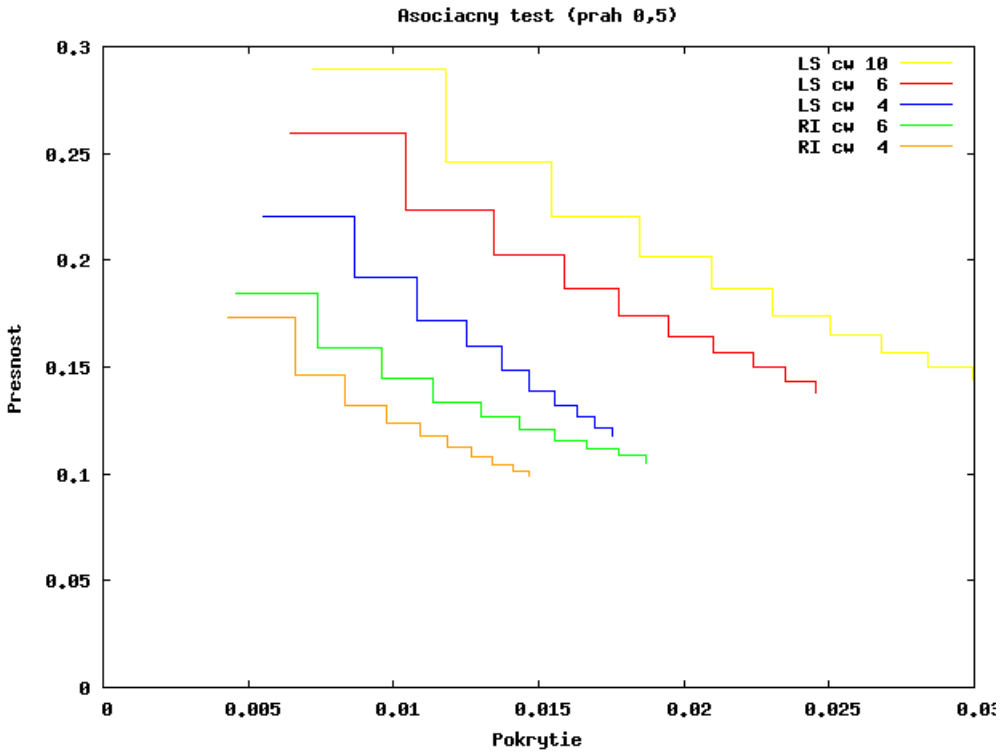
Tabuľka 7.2 Hodnoty pokrytia a presnosti pri zadanom prahu 0,5 a veľkosti kontextového okna 4

Podobné termíny	Random Indexing		Lexikálna substitúcia	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**
1	0,004597	0,184864	0,006456	0,259682
2	0,007412	0,158903	0,010424	0,223514
3	0,009605	0,144475	0,013439	0,202165
4	0,011342	0,133475	0,015856	0,186617
5	0,012993	0,127018	0,017781	0,173808
6	0,014328	0,120749	0,019489	0,164212
7	0,015532	0,115749	0,021030	0,156663
8	0,016671	0,111754	0,022408	0,150132
9	0,017765	0,108609	0,023472	0,143406
10	0,018674	0,105172	0,024562	0,138229
Všetky	0,069074	0,021862	0,069084	0,021846

Tabuľka 7.3 Hodnoty pokrytia a presnosti pri zadanom prahu 0,5 a veľkosti kontextového okna 6

Podobné termíny	Random Indexing		Lexikálna substitúcia	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**
1	0,004641	0,185899	0,007224	0,289309
2	0,007721	0,160550	0,011818	0,245760
3	0,010262	0,146271	0,015442	0,220140
4	0,012472	0,136353	0,018481	0,202105
5	0,014387	0,128310	0,020968	0,187078
6	0,016037	0,121292	0,023026	0,174243
7	0,017422	0,114760	0,025041	0,165027
8	0,018874	0,110429	0,026792	0,156840
9	0,020232	0,106696	0,028388	0,149784
10	0,021507	0,103389	0,029925	0,143938
Všetky	0,061116	0,042049	0,061049	0,042070

Tabuľka 7.4 Hodnoty pokrytia a presnosti pri zadanom prahu 0,5 a veľkosti kontextového okna 10



Obrázok 7.3 Graf závislosti presnosti od pokrytia výsledkov systému voči asociačnému testu pre experiment s termínmi podobnejšími než 0,5

V grafe 7.3 sú zaznamenané hodnoty z tabuliek 7.2 a 7.3. Z tabuľky 7.4 je v grafe zobrazená iba časť pre lexikálnu substitúciu. Hodnoty pre Random Indexing sú pri veľkosti kontextového okna 10 skoro rovnaké ako pri veľkosti 6, takže ich zobrazenie by mohlo znížiť čitateľnosť grafu. Všetky krivky v grafe zaznamenávajú závislosť presnosti od pokrytia pre metódy Random Indexing a lexikálna substitúcia pre veľkosti kontextového okna 4 a 6 a v prípade lexikálnej substitúcie aj pre veľkosť kontextového okna 10.

Je vidieť, že výsledky pre lexikálnu substitúciu sú vždy lepšie ako výsledky pre Random Indexing, takže lexikálna substitúcia dokáže usporiadať kandidátne termíny získané pomocou metódy Random Indexing tak, že podobné termíny dá viac dopredu. Takže aj výsledky v asociačnom teste sú lepšie. Ďalej je vidieť, že pri veľkosti kontextového okna 6 dáva metóda Random Indexing lepšie výsledky ako pri veľkosti 4, aj keď tieto výsledky nie sú výrazne lepšie. Pri zväčšení okna hodnotu 10 sú výsledky prakticky rovnaké ako pri veľkosti 6. Pri lexikálnej substitúcii je rozdiel medzi výsledkami pre veľkosť kontextového okna 4, 6 a 10 viac výrazný, pričom najlepšie sú výsledky pre veľkosť 10 a najhoršie pre veľkosť 4. Pri veľkosti 10 sa hodnota presnosti blíži maximálnej hodnote presnosti získanej v [9], ktorá dosahovala hodnoty 0,35, ako je možné vidieť v grafe 7.1. Hodnoty pokrytia sú však výrazne nižšie (v [9] bolo pokrytie pri porovnateľnej presnosti 0,05, čiže 10krát väčšie).

Z výsledkov systému sa dá usudzovať, že keď sa pri metóde Random Indexing berie do úvahy väčší kontext termínov, tak systém je schopný lepšie určiť kandidátne termíny a aj pri malom zlepšení kandidátnych termínov metóda lexikálnej substitúcie výraznejšie zlepší výsledky. Na tomto zlepšení

sa môže podieľať aj to, že pri väčšom kontextovom okne sa skúma viac termínov. Takže sa skúmajú aj termíny, pre ktoré sa našli podobné termíny, ale pri menšej veľkosti kontextového okna podobnosť týchto kandidátov so skúmaným termínom neprekročila hodnotu 0,5, a teda tieto termíny sa následne nebrali do úvahy.

Pri porovnaní výsledkov s výsledkami získanými porovnávaním s WordNetom je vidieť, že pri WordNete vyšli vyššie hodnoty pokrytia ale nižšie hodnoty presnosti. Nižšie hodnoty pokrytia sú spôsobené tým, že asociačný test obsahuje viac podobných termínov k jednotlivým skúmaným termínom ako WordNet (niekoľko desiatok v asociačnom teste oproti menej ako 10 vo WordNete). Takže pri počítaní pokrytia, keď sa berie do úvahy maximálne 10 najpodobnejších termínov, je hodnota *fn* (vysvetlené vo vzorci 6.3) oveľa vyššia pri asociačnom teste ako pri WordNete, takže pokrytie vyjde nižšie. Naopak presnosť vyšla lepšie pri asociačnom teste. Rolu tu hrá opäť aj to, že asociačný test obsahuje viac podobných termín ku každému skúmanému termínu, takže pri porovnávaní je väčšia pravdepodobnosť zhody. Asociačný test však obsahuje aj iné slová ako synonymá. A ako bolo ukázané v kapitole 7.1, systém vyberá často ako podobné termíny nie synonymá, ale slová, ktoré patria do rovnakej domény.

Nasledujúce tabuľky a graf zobrazujú hodnoty presnosti a pokrytia pre experiment, kedy sa bralo pre každý skúmaný termín do úvahy prvých 100 najpodobnejších termínov, ktoré boli určené metódou Random Indexing, bez ohľadu na to, ako podobné boli s daným skúmaným termínom. Veľkosť kontextového okna bola opäť 4, 6 a 10. 100 najpodobnejších termínov tvorí menej ako 0,5% všetkých možných termínov, ktoré sa vyskytli ako asociácie v asociačnom teste. Pri tomto experimente bolo použitých 7 142 termínov z asociačného testu, ktoré mal systém uložené v indexe.

Podobné termíny	**Random Indexing**		**Lexikálna substitúcia**	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**
1	0,003498	0,139012	0,005905	0,249627
2	0,005868	0,116582	0,010845	0,226410
3	0,007861	0,104148	0,015096	0,208400
4	0,009750	0,096892	0,018431	0,189778
5	0,011394	0,090601	0,021188	0,173720
6	0,013010	0,086213	0,023702	0,161335
7	0,014420	0,081906	0,026103	0,151861
8	0,015740	0,078230	0,028292	0,143724
9	0,017112	0,075603	0,030415	0,137076
10	0,018424	0,073273	0,032308	0,130871
Všetky	0,076680	0,030531	0,076679	0,030527

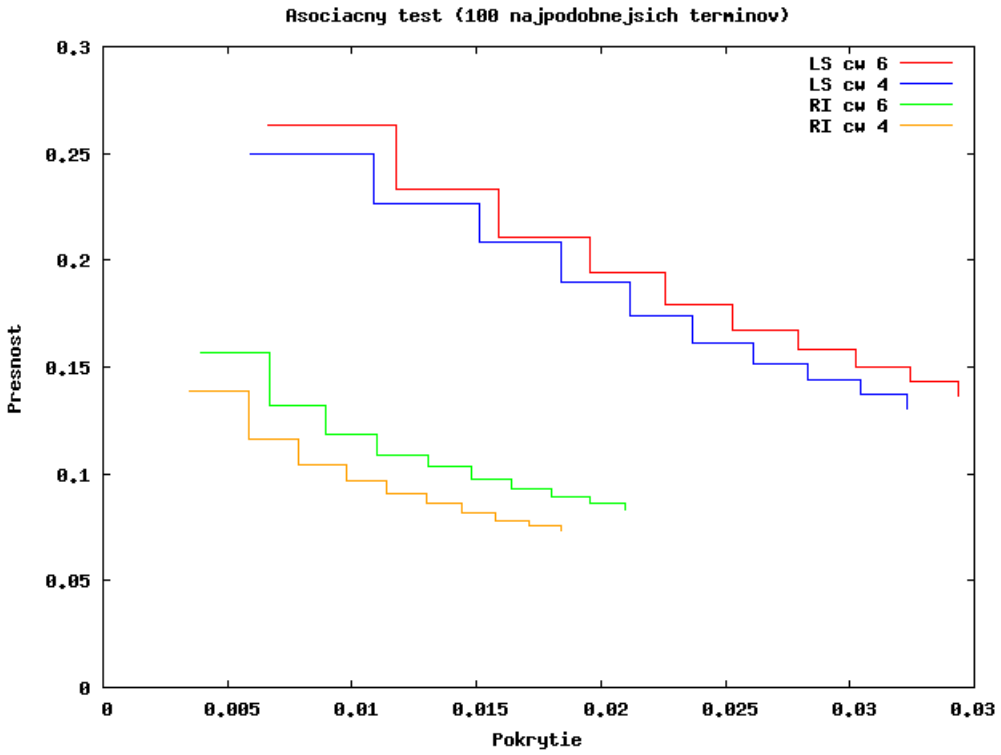
Tabuľka 7.5 Hodnoty pokrytia a presnosti pri prvých 100 termínoch pre veľkosť kontextového okna 4

Podobné termíny	Random Indexing		Lexikálna substitúcia	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**
1	0,003953	0,156939	0,00663	0,263232
2	0,006659	0,132192	0,011744	0,233128
3	0,008932	0,118212	0,015905	0,210492
4	0,010982	0,109005	0,019551	0,194063
5	0,013035	0,103508	0,022598	0,179446
6	0,014788	0,097856	0,025310	0,167483
7	0,016391	0,092967	0,027927	0,158399
8	0,018015	0,089405	0,030233	0,150046
9	0,019561	0,086295	0,032395	0,142910
10	0,021004	0,083391	0,034384	0,136516
Všetky	0,083864	0,033297	0,083908	0,033317

Tabuľka 7.6 Hodnoty pokrytia a presnosti pri prvých 100 termínoch pre veľkosť kontextového okna 6

Podobné termíny	Random Indexing		Lexikálna substitúcia	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**
1	0,004113	0,163316	0,006633	0,263372
2	0,007089	0,140718	0,011807	0,234388
3	0,009641	0,127594	0,016014	0,211939
4	0,011946	0,118571	0,019774	0,196269
5	0,013988	0,111077	0,022764	0,180762
6	0,015871	0,105019	0,025356	0,167787
7	0,017498	0,099245	0,027969	0,158639
8	0,019177	0,095175	0,030311	0,150431
9	0,020808	0,091794	0,032519	0,143455
10	0,022289	0,088493	0,034575	0,137272
Všetky	0,085307	0,033869	0,085364	0,033895

Tabuľka 7.7 Hodnoty pokrytia a presnosti pri prvých 100 termínoch pre veľkosť kontextového okna 10



Obrázok 7.4 Graf závislosti presnosti od pokrytia výsledkov systému voči asociačnému testu pre experiment s najpodobnejšími 100 termínmi

V grafe 7.4 sú zaznamenané hodnoty z tabuliek 7.5 a 7.6. Nie sú v ňom hodnoty z tabuľky 7.7, pretože tieto hodnoty sú skoro rovnaké ako v tabuľke 7.6, čo by znížilo čitateľnosť grafu. Všetky krivky v grafe zaznamenávajú závislosť presnosti od pokrytia pre metódy Random Indexing a lexikálna substitúcia pre veľkosti kontextového okna 4 a 6.

Z výsledkov je opäť vidieť, že výsledky pre lexikálnu substitúciu sú lepšie ako výsledky pre Random Indexing. Rovnako výsledky pre veľkosť kontextového okna 6 sú lepšie ako pre veľkosť 4 a skoro rovnaké ako pre veľkosť 10. V tomto prípade metóda lexikálnej substitúcie vylepšila výsledky Random Indexingu výraznejšie ako v prípade, že sa ako kandidátne termíny brali tie, ktoré mali so skúmaným termínom podobnosť aspoň 0,5. To je pravdepodobne spôsobené tým, že Random Indexing poskytol v tomto prípade ako kandidátne termíny aj tie, ktoré mali nižšiu hodnotu podobnosti. Bolo teda poskytnutých viac kandidátnych termínov, takže lexikálna substitúcia mala väčšiu možnosť výberu medzi kandidátnymi termínmi, a teda bola aj väčšia pravdepodobnosť, že medzi kandidátnymi termínmi sú aj skutočne podobné termíny. Zväčšenie veľkosti kontextového okna z hodnoty 4 na hodnotu 6 už prílišné zlepšenie výsledkov neprinieslo a zväčšenie zo 6 na 10 neprinieslo takmer žiadne zlepšenie. Pravdepodobne je to preto, lebo medzi prvými 100 najpodobnejšími termíny neprišlo až k toľkým zmenám. Pri experimente, keď sa použila prahová hodnota podobnosti 0,5 došlo k väčším zmenám vo výsledkoch aj preto, lebo pri rozdielnych veľkostiach kontextového okna dochádzalo k pridávaniu alebo odoberaniu kandidátnych termínov. Pri experimente, keď sa bralo do úvahy 100 najpodobnejších termínov k takýmto zmenám už

nedochádzalo, pretože všetky termíny, ktoré sa mohli pridať alebo odobrať kvôli prekročeniu alebo neprekročeniu prahu sú obsiahnuté v týchto 100 termínoch, a teda zmeny nastávajú skôr v poradí kandidátnych termínov, čo nemá pri metóde lexikálnej substitúcie žiadny vplyv na výsledky.

Celkovo je na výsledkov vidieť, že pokrytie je veľmi malé. Toto je spôsobené tým, že sa skúma len niekoľko málo najpodobnejších termínov. Je vidieť, že so vzrastajúcim počtom skúmaných podobných termínov rastie pokrytie, výrazne však klesá presnosť. Čím menej najpodobnejších termínov sa skúma, tým sú výsledky presnejšie. Aj napriek nárastu pokrytia so stúpajúcim počtom najpodobnejších termínov, nepresahuje hodnota pokrytia 0,1, takže systém nedokáže nájsť veľké množstvo termínov, ktoré sú považované ako podobné v asociačnom teste. Pri porovnaní s výsledkami z [9] dosahuje systém porovnateľnú presnosť (0,25 – 0,28 oproti 0,35), ale pokrytie je výrazne nižšie (0,005 oproti 0,05).

Presnosť systému dosahuje maximálne 0,25, čo nie je veľa. Vplýva na to viacero faktorov. Prvým je ten, že systém môže v niektorých prípadoch vyhľadávať podobné termíny iným spôsobom ako ľudia. Napríklad systém vybral pre termín „injection“ ako najpodobnejší termín „inject“. Slová spolu súvisia, avšak jedná sa o rôzne slovné druhy. V asociačnom teste sú ako podobné termíny pre „injection“ uvedené napríklad „needle“, „pain“ alebo „jab prick“.

Ďalším faktorom je, že systém niekedy priradí ako najpodobnejšie slovo k termínu množné číslo tohto termínu, prípadne jednotné číslo, pokiaľ bol pôvodný termín v množnom čísle. Takéto prípady tiež nie sú zahrnuté v asociačnom teste. V systéme by sa nemali vyskytovať termíny v množných číslach, pretože sa používajú základné tvary slov. Ich výskyt je však spôsobený tým, že v SW1 korpuse môžu byť niekedy zle určené základné tvary slov a tak sa môže stať, že sa v systéme vyskytne termín v jednotnom aj v množnom čísle. Tento prípad však vo výsledkoch nie je príliš častý.

Ďalším faktorom, ktorý má vplyv na výsledky je viacznačnosť termínov. Napríklad pre termín „jam“ sú v asociačnom teste uvedené ako podobné termíny „strawberry“, „bread“, „marmalade“. Systém určil ako podobné termíny „music“, „bob dylan“, „rock“. Takže je vidieť, že systém vybral podobné termíny pre iný význam termínu „jam“ ako je v asociačnom teste. Vo Wikipédii sa totiž slovo „jam“ objavuje v názvoch hudobných skupín („The Jam“, „Pearl Jam“). Zlepšiť výsledky pre viacvýznamové slová by malo porovnávanie konceptov s termínmi, avšak pre asociačný test by bolo ťažké nájsť koncepty pre termíny. Napríklad pre termín „jam“ vo význame „želé“, čo je význam použitý v asociačnom teste, je na Wikipédii článok s názvom „Fruit preserves“. Takže by bolo potrebné všetky termíny mapovať na koncepty ručne, čo je vzhľadom na počet termínov v asociačnom teste (viac ako 8000) časovo extrémne náročné. Navyše by sa pre veľký počet termínov pravdepodobne nenašiel zodpovedajúci koncept.

Pri experimentoch s asociačným testom vyšli vyššie hodnoty presnosti oproti experimentom s WordNetom, pokrytie však vyšlo nižšie. Je to tým, že v asociačnom teste je pre každý skúmaný termín viac asociácií, ako je synoným vo WordNete. Navyše v asociačnom teste nie sú len synonymá, ale aj iné podobné termíny.

7.3 WordSimilarity-353

Pri tomto experimente sa vybrali všetky termíny, ktoré sú v sade WordSimilarity-353 a porovnali sa každý s každým. Celkovo je v sade 437 termínov (353 párov, niektoré termíny sa opakujú). Pre každý termín boli ako kandidátne termíny vybrané všetky ostatné termíny z tejto sady a boli usporiadané metódou Random Indexing. Následne bolo toto poradie ešte preusporiadané a hodnoty podobnosti boli spočítané metódou lexikálnej substitúcie. Počítanie podobnosti každého termínu s každým iným pri metóde Random Indexing je použité preto, lebo systém je navrhnutý pre tvorbu tezauru, takže pre každý termín vyberá zo zadaného zoznamu niekoľko najpodobnejších termínov a bolo by pomerne komplikované porovnávať iba konkrétne dvojice termínov. Na výsledkov však tento postup nemá vplyv, pretože vo výsledkoch je porovnaný každý termín s každým, takže sú porovnané všetky dvojice termínov zo sady WordSimilarity-353.

Korelácia konečných výsledkov systému so sadou WordSimilarity-353 vyšla 0,19279. Korelácia výsledkov Random Indexingu s touto sadou bola 0,26809. Pre porovnanie sa ako ohodnotenie podobných termínov nepoužili hodnoty, ktoré určil systém, ale hodnoty odvodené od poradia termínov, takže najpodobnejší termín bude mať ohodnotenie 437, každým ďalším termínom sa toto ohodnotenie zníži o 1 až posledný (437.) bude mať ohodnotenie 1. V takomto prípade vyšla korelácia konečných výsledkov 0,28399 a korelácia výsledkov Ranom Indexingu 0,27161.

Ako je vidieť, korelácia nebola príliš veľká. Existujú metódy, ktoré dosahujú vyššiu koreláciu, ako je možné vidieť v tabuľke 6.1. Ohodnotenie, ktoré termínom priradila lexikálna substitúcia, má veľmi nízku koreláciu. Je to spôsobené najmä tým, že tieto hodnoty majú veľký rozptyl, keď sa ohodnotenie najpodobnejších termínov pohybuje od hodnôt menších ako 1 po hodnoty rádovo v tisícoch. Keď sa použili namiesto týchto hodnôt, hodnoty získané z usporiadania termínov, korelácia bola vyššia. Korelácia výsledkov systému síce nie je vysoká, ale približuje sa výsledkom získaným pomocou WordNetu, ktorý bol tiež vytvorený ručne.

7.4 Vyhodnotenie výsledkov pre koncepty

Experiment pre vyhľadávanie podobných termínov pre koncepty prebiehal tak, že sa pre koncepty našlo 1 000 najpodobnejších termínov pomocou metódy Random Indexing a tie sa následne preusporiadali metódou lexikálnej substitúcie. 1 000 termínov je menej ako 0,1% zaindexovaných termínov. Na vytvorenie kontextových vektorov pre koncepty boli použité dve metódy, a to vytvorenie vektorov z kontextových vektorov termínov, ktoré daný koncept obsahoval, a vytvorenie vektorov z náhodných vektorov pre tieto termíny. V ďalšom experimente sa metódou Random Indexing našlo 100 najpodobnejších konceptov k zadaným konceptom (porovnávali sa teda v podstate články Wikipédie metódou Random Indexing). V prípade, že sa vyberali najpodobnejšie termíny boli hodnoty presnosti aj pokrytia pre výsledky získané metódou Random Indexingu vždy rovné 0.

Random Indexing teda nedokázal ako najpodobnejšie vybrať zodpovedajúce termíny. Lexikálna substitúcia následne dokázala kandidátne termíny usporiadať tak, že ako najpodobnejšie vybrala aspoň nejaké skutočne podobné termíny. V prípade, že sa porovnávali koncepty s konceptmi, sa nedala použiť lexikálna substitúcia, pretože veľa nájdených konceptov bolo dlhších ako päť slov alebo obsahovali v zátvorke doplňujúce informácie. Na takéto prípady nie je možné použiť lexikálnu substitúciu, lebo pre takéto termíny by sa nenašli žiadne kontexty vo Web 1T korpuse. Random Indexing však dokázal zo všetkých konceptov v indexe vybrať ako najpodobnejšie aj také koncepty, ktoré sú so zadanými konceptmi podobné aj v skutočnosti.

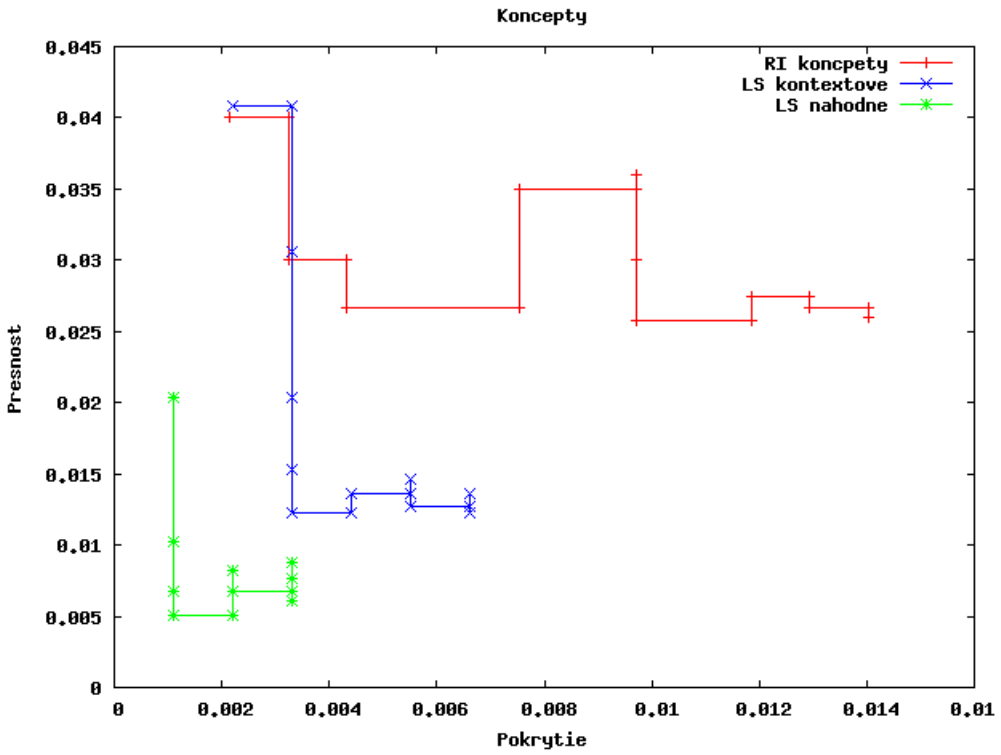
Výsledky systému sa porovnávali s termínmi, ktoré boli uvedené v tezaure[10], pričom sa brali do úvahy iba synonymá k tým významom konceptu, ktorý najlepšie zodpovedal obsahu článku na Wikipédii, ktorý tento koncept popisoval.

V tabuľke 7.8 a grafe 7.5 sú hodnoty presnosti a pokrytia pre lexikálnu substitúciu, keď sa na výpočet kontextových vektorov pre koncepty použili kontextové vektory pre termíny (LS kontextové), keď sa použili náhodné vektory pre termíny (LS náhodné) a keď sa metódou Random Indexing porovnávali koncepty s konceptmi (RI koncepty). Nie sú uvedené hodnoty pre Random Indexing pri porovnávaní konceptov s termínmi, pretože všetky hodnoty boli nulové.

Podobné termíny	**LS kontextové**		**LS náhodné**		**RI koncepty**	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**	**pokrytie**	**presnosť**
1	0,002205	0,040816	0,001103	0,020408	0,002157	0,040000
2	0,003308	0,030612	0,001103	0,010204	0,003236	0,030000
3	0,003308	0,020408	0,001103	0,006803	0,004315	0,026667
4	0,003308	0,015306	0,001103	0,005102	0,007551	0,035000
5	0,003308	0,012245	0,002205	0,008163	0,009709	0,036000
6	0,004410	0,013605	0,002205	0,006803	0,009709	0,030000
7	0,005513	0,014577	0,003308	0,008746	0,009709	0,025714
8	0,005513	0,012755	0,003308	0,007653	0,011866	0,027500
9	0,006615	0,013605	0,003308	0,006803	0,012945	0,026667
10	0,006615	0,012245	0,003308	0,006122	0,014024	0,026000
Všetky	0,054024	0,001047	0,000671	0,012128	0,025890	0,004800

Tabuľka 7.8 Hodnoty presnosti a pokrytia pre experiment s koncpetmi

[10] tezaurus.com



Obrázok 7.5 Závislosť presnosti od pokrytia výsledkov systému voči tezauru pre experiment s konceptmi

V grafe 7.5 je vidieť, že závislosť presnosti od pokrytia nie je klesajúca. To je spôsobené tým, že bolo skúmaných iba 50 konceptov. Tiež bolo nájdených málo termínov zhodných so zadanou sadou, ktorá bola vytvorená z existujúceho tezauru. Keď sa našiel nejaký zhodný podobný termín, spôsobilo to skokové zvýšenie presnosti. Niekedy sa pri zvýšení počtu najpodobnejších termínov, ktoré sa brali do úvahy nenašla žiadna ďalšia zhoda. V takom prípade ostala hodnota pokrytia rovnaká, ale hodnota presnosti klesla.

Z výsledkov je vidieť, že pri experimentoch s konceptmi boli hodnoty presnosti a pokrytia veľmi nízke. Najlepšie dopadli experimenty, keď sa porovnávali koncepty s konceptmi. Spôsobené je to najmä tým, že kontextové vektory pre koncepty sa nedajú dobre porovnať s kontextovými vektormi pre termíny. Systém potom nedokáže vhodne vybrať kandidátne termíny. Pri porovnávaní konceptov s konceptmi dochádza k zlepšeniu výsledkov. Problémom je, že v systéme nie je dosť konceptov, ktoré by sa dali považovať za podobné k zadaným konceptom, pretože Wikipédia takéto články neobsahuje.

V grafe 7.1 sú pre porovnanie uvedené výsledky z [9], kde sa vyhodnocovala presnosť a pokrytie rovnakých konceptov voči synonymám získaným z Oxford American Writer's Thesaurus. Ako je vidieť výsledky tohto systému sú lepšie. V tomto prípade však boli lexikálnej substitúcii predané kvalitnejšie kandidátne termíny (krivka Baseline). Systém popísaný v tejto práci nedokázal poskytnúť také kvalitné kandidátne termíny pre koncepty a preto sú aj celkové výsledky oveľa horšie.

V tabuľke 7.9 je prvých 5 najpodobnejších termínov, respektíve konceptov, k niektorým konceptom z tabuľky 6.2. Veľkým začiatočným písmenom sú označené koncepty a malým termíny.

Koncept	LS kontextové	LS náhodné	RI koncepty
Insomnia	at least	medical equipment	Narcolepsy
	loss	wireless communication	Sleep deprivation
	poor	establishment	Major depressive episode
	difficulty	abilities	Anxiety
	at least two	theatre company	Ménière's disease
Computer	at least	software	Computer/Temp
	system	hardware	Personal computer
	business	video game	Timeline of computing 1950-1979
	small	websites	IBM System/360
	program	data management	Reconfigurable computing
Earth	earth	atmosphere	Space colonization
	at least	establishment	Venus
	life	orbit	Titan (moon)
	body	geology	List of misconceptions
	system	data management	Mercury (planet)
God	god	human	Monotheism
	at least	faith	Pantheism
	man	word	God and gender
	life	divine	Predestination
	christian	satan	Abrahamic religion

Tabuľka 7.9 Najpodobnejších 5 termínov a konceptov pre vybrané koncepty

Z tabuľky je vidieť, že najlepšie podobné termíny boli vybrané, keď sa porovnávali koncepty s konceptmi. Nejedná sa síce priamo o synonymá, vybrané koncepty však patria do rovnakej domény, ako zadaný koncept alebo so zadaným konceptom priamo súvisia (God - Monotheism). Pri porovnávaní konceptov s termínmi sú výsledky horšie. Často sa tu vyskytujú všeobecné slová, ktoré nemajú so zadaným konceptom nič spoločné („at least“).

Pri niektorých konceptoch sa ako najpodobnejšie termíny vybrali tie isté slová (napríklad pre koncept „God“ sa vybral ako najpodobnejší termín „god“). To je preto, lebo sa porovnáva koncept s termínom, a teda sa nejedná o to isté. Pri porovnávaní termínov s termínmi alebo konceptov s konceptmi sa termín alebo koncept neporovnáva sám so sebou. Ak sa pre koncept ako najpodobnejší termín vybralo rovnaké slovo, tento termín sa ignoruje. Keby sa bral do úvahy ako podobný termín, tak by presnosť a pokrytie dosahovali hodnôt uvedených v tabuľke 7.10. Takýto postup sa však v praxi nepoužíva, pretože pri tvorbe tezauru je nežiaduce mať medzi podobnými termínmi to isté slovo, ku ktorému sa hľadajú podobné slová.

Podobné termíny	LS kontextové		LS náhodné	
	pokrytie	**presnosť**	**pokrytie**	**presnosť**
1	0,017641	0,326531	0,004410	0,081633
2	0,020948	0,193878	0,004410	0,040816
3	0,020948	0,129252	0,004410	0,027211
4	0,020948	0,096939	0,004410	0,020408
5	0,020948	0,077551	0,005513	0,020408
6	0,022051	0,068027	0,005513	0,017007
7	0,023153	0,061224	0,006615	0,017493
8	0,023153	0,053571	0,006615	0,015306
9	0,024256	0,049887	0,006615	0,013605
10	0,024256	0,044898	0,006615	0,012245
Všetky	0,071665	0,001389	0,015436	0,000855

Tabuľka 7.10 Presnosť a pokrytie výsledkov systému pre koncepty voči existujúcemu tezauru, ak sa brali do úvahy aj zhodné termíny

7.5 Časová náročnosť systému

Na získanie doby behu sa použil nástroj *time*[11]. Hodnotila sa časová náročnosť jednotlivých častí systému. Doba trvania extrakcie termínov je vždy rovnaká. Aj doba indexácie je pri každom behu systému rovnaká. Ich trvanie síce závisí na veľkosti vstupu, systém však používa SW1 korpus, ktorého veľkosť sa nemení. Vytváranie kontextových vektorov závisí od dimenzie vektorov a veľkosti kontextového okna. V systéme sa pre dimenziu vektorov používa hodnota 1 000, ktorá sa ukázala ako dostačujúca pre fungovanie systému. Závislosť času vytvorenia vektorov od veľkosti kontextového okna nie je príliš veľká, pretože ovplyvňuje iba počet pripočítaní náhodného vektoru ku kontextovému, čo zaberá málo času v porovnaní s inými operáciami, ako je napríklad načítanie dokumentov z indexu. Vyhľadávanie kandidátnych termínov závisí od počtu skúmaných a potenciálnych kandidátnych vektorov. Presnejšie je priamo úmerná súčinu týchto veľkostí. Extrakcia kontextov pre skúmané termíny je závislá od počtu skúmaných termínov, nie však príliš výrazne. Kontexty sa totiž vyhľadávajú pri jednom sekvenčnom prechode SW1 korpusu a počet skúmaných termínov nemá až takú váhu. Vyhľadávanie pravých a ľavých kontextov závisí od počtu týchto kontextov. Presnejšie závisí od počtu súborov, v ktorých sú tieto kontexty uložené. Každý súbor bude predstavovať jeden sekvenčný priechod časťou (podľa dĺžky kontextov) Web 1T korpusu. Vyhľadávanie kontextov pre kandidátne termíny závisí od počtu týchto termínov. Vždy vzniknú 4 súbory pre n-gramy (2<=n<=5) a pre každý z týchto súborov sa vykoná jeden sekvenčný priechod

[11] http://unixhelp.ed.ac.uk/CGI/man-cgi?time

príslušnou časťou (podľa dĺžky kontextov) Web 1T korpusu. Počet termínov ovplyvňuje veľkosť týchto súborov a tá ovplyvňuje rýchlosť porovnávania každého záznamu z Web 1T korpusu s obsahom tohto súboru. Filtrovanie a ohodnotenie kontextov pre kandidátne termíny je závislé od počtu nájdených kontextov, ktorý je závislí od počtu kandidátnych termínov. Hľadanie podobných termínov metódou lexikálnej substitúcie je závisle od počtu skúmaných a kandidátnych termínov. Presnejšie je závislé od počtu porovnaní skúmaných a kandidátnych termínov.

Časť systému		**Doba trvania**
Extrakcia termínov		1 h 30 m
Indexácia		1 h 30 m
Vytváranie kontextových vektorov	veľkosť kontextového okna 4	6 h 30 m
	veľkosť kontextového okna 10	6 h 40 m
Počítanie podobnosti termínov metódou Random Indexing	8 477 253 184 porovnaní	11 h 20 m
	177 241 porovnaní	0 h 25 m
	275 292 064 porovnaní	2 h 20 m
Extrakcia kontextov pre skúmané termíny	92 072 termínov	3 h 0 m
	353 termínov	2 h 0 m
Vyhľadávanie pravých a ľavých kontextov vo Web 1T korpuse	92 072 termínov[12]	9 h 20 m
	353 termínov[13]	4 h 30 m
Vyhľadávanie kontextov pre kandidátne termíny vo Web 1T korpuse	92 072 termínov	4 h 30 m
	353 termínov	1 h 30 m
Filtrovanie a ohodnotenie kontextov pre kandidátne termíny	92 072 termínov	19 h 45 m
	353 termínov	1 h 0 m
Hľadanie podobných termínov metódou lexikálnej substitúcie	1 136 748 porovnaní	49 h 40 m
	177 241 porovnaní	3 h 20 m
	325 627 porovnaní	11 h 30 m

Tabuľka 7.11 Časová náročnosť jednotlivých častí systému

[12] celkovo bolo vytvorených 6 súborov s pravými a ľavými kontextami (2 pre dĺžku 2 slová, 2 pre dĺžku 3 a 2 pre dĺžku 4)

[13] celkovo bolo vytvorených 5 súborov s pravými a ľavými kontextami (1 pre dĺžku 2 slová, 2 pre dĺžku 3 a 2 pre dĺžku 4)

8 Záver

Cieľom tejto práce bolo vytvoriť systém, ktorý je schopný vytvoriť tezaurus s použitím Wikipédie. K tomu bola použitá metóda Random Indexing na generovanie kandidátnych termínov a metóda lexikálnej substitúcie na preusporiadanie kandidátnych termínov. Pre implementáciu systému sa použili knižnice Apache Lucene, semanticvectors a nástroj Get1T. Pri výpočte podobnosti termínov čerpal systém zdrojové údaje zo SW1 korpusu a Web 1T korpusu.

Náročnou úlohou bolo vyhodnotenie výsledkov, ktoré systém poskytol. Počítala sa korelácia so sadou WordSimilarity-353. Táto hodnota nevyšla príliš veľká. Blížila sa však hodnote korelácie podobnosti získanej z WordNetu, ktorý bol tiež vytvorený ručne, s touto sadou, takže táto metrika sa nedá považovať za veľmi dôležitú. Navyše v tezauroch nebývajú hodnoty podobnosti jednotlivých termínov.

Pri tvorbe tezauru je oveľa dôležitejšie usporiadanie podobných termínov tak, aby sa medzi najpodobnejšími termíny vyskytlo čo najviac termínov, ktoré sú so zadaným termínom podobné aj v skutočnosti. Systém je schopný usporiadať termíny takýmto spôsobom. Toto sa ukázalo aj pri porovnávaní s asociačným testom a WordNetom, keď najlepšia presnosť výsledkov bola dosiahnutá, keď sa bral do úvahy iba jeden najpodobnejší termín pre každý skúmaný termín. Hodnoty presnosti potom postupne klesali pre prvé dva, tri až desať najpodobnejších termínov, takže systém usporiadal termíny tak, že sa snažil dať dopredu tie termíny, ktoré sú podobné aj v skutočnosti. Systém však nenašiel veľké množstvo podobných termínov uvedených v asociačnom teste, čo dokazuje aj nízka hodnota pokrytia. Našiel však iné termíny, ktoré by sa dali považovať za podobné, prípadne podobné termíny pre iný význam termínu, než sa použil v asociačnom teste. V prípade porovnávania s WordNetom bolo pokrytie vyššie ako pri asociačnom teste, nižšia však bola presnosť. Spôsobené je to najmä tým, že asociačný test obsahuje viac podobných termínov pre jednotlivé skúmané termíny. Tieto podobné termíny často nie sú synonymá, ale slová z rovnakej domény. Pri porovnaní zistených hodnôt presnosti a pokrytia s výsledkami z [9] bolo zistené, že hodnoty presnosti boli porovnateľné (maximálne hodnoty boli 0,28 oproti 0,35), ale hodnoty pokrytia boli oveľa nižšie (0,005 až 0,18 oproti 0,05 až 0,35). Porovnateľná presnosť bola dosiahnutá aj napriek tomu, že v [9] porovnávali koncepty s termínmi na rozdiel od systému prezentovanému v tejto práci, ktorý počítal podobnosť termínov s termínmi. Mnohé z týchto termínov majú viac významov, čo má vplyv na výsledky.

Z výsledkov je vidieť, že lexikálna substitúcia je veľmi závislá od kandidátnych termínov poskytnutých metódou Random Indexing. Ich kvalita veľmi ovplyvňuje aj kvalitu celkového výstupu systému, čo je možné vidieť pri experimentoch s konceptmi. Ich počet zas ovplyvňuje dobu behu systému, pretože lexikálna substitúcia je časovo náročná metóda a systém by sa mal snažiť o to, aby neposkytoval príliš veľký počet kandidátnych termínov.

Úspešnosť systému pri vyhľadávaní synoným a iných podobných termínov je obmedzená aj nedeterminizmom prirodzeného jazyka, ktorý spôsobujú napríklad viacvýznamové slová. S týmto malo pomôcť porovnávanie konceptov s termínmi. Týmto spôsobom boli dosiahnuté veľmi dobré výsledky v [9]. Kvôli ťažkostiam so spôsobom reprezentácie konceptov však systém nedosahoval v tomto prípade dobré výsledky. Bolo to spôsobené najmä tým, že pri metóde Random Indexing nebolo možné vhodným spôsobom porovnať koncepty s termínmi a vygenerované kandidátne termíny sa ukázali ako nevhodné. Pri porovnávaní konceptov s konceptmi sa ukázalo, že nájdené podobné koncepty skutočne súvisia so zadaným konceptom. Nastal však problém s tým, že v indexe nebolo dosť vhodných konceptov. Každý koncept totiž predstavoval článok vo Wikipédii a medzi článkami sa obvykle nenachádzajú dva také, ktoré by popisovali termíny, ktoré sú synonymá. Navyše nie vždy sa musí vo Wikipédii nachádzať článok popisujúci koncept, ku ktorému chceme nájsť podobné termíny. V takom prípade sa s konceptmi nedá pracovať a je potrebné počítať podobnosť termínov s termínmi.

Úspešnosť systému v testoch znižovalo aj to, že podobnosť termínov nie je jasne definovaná a systém vybral ako najpodobnejšie také termíny, ktoré sa dajú považovať za blízke ku skúmanému termínu, ale porovnávané sady ich neobsahovali. Systém nevyberal vždy ako najpodobnejšie termíny synonymá, ale väčšinou vybral termíny, ktoré patrili do rovnakej domény. Napríklad k dopravným prostriedkom vybral iné dopravné prostriedky, k chorobám iné choroby a podobne. Horšie výsledky boli dosahované pri všeobecných termínoch („wind“, ...) ako pri viac špecifických (názvy dopravných prostriedkov, mená zvierat, rastlín, ...).

Systém je možné vylepšiť niekoľkými smermi. Prvým je zlepšenie práce s konceptmi. Konkrétne zlepšenie reprezentácie konceptov a ich porovnávanie s termínmi. V tomto prípade je možnosť zvoliť inú formu generovania kandidátnych termínov pre kontexty. Ako príklad môže poslúžiť [9], kde sa ku konceptom vyberali ako kandidátne termíny tie, ktoré sa vyskytli v hypertextových odkazoch na články, ktoré popisovali daný koncept. Je možné zvoliť aj inú formu oddeľovania rôznych významov slov, napríklad pre každý význam slova zvoliť inú textovú reprezentáciu. Automaticky rozlišovať významy slov je však veľmi náročné. Ďalšou možnosťou vylepšenia je zlepšenie extrakcie termínov, pretože aj kvalita vybraných termínov má vplyv na výsledky systému.

Literatúra

[1] Backfield, G., Otrusina, L., Smrž, P.: M-Eco D3.1 – Speech Recognition and Content Classification Subsystems, Hannover, DE, EU-7FP-ICT, 2010

[2] Bernard, J. N. L. (Editor): The Macquarie Thesaurus, 2nd Ed, 2007 ISBN-13**:** 9781876429614

[3] Brand, M.: Fast Low-Rank Modifications of the Think Singular Value Decomposition, Linear Algebra and Its Applications, Vol. 415, Issue 1, May 2006

[4] Denhière G., Lemaire B.: Representing children's semantic knowledge from a multisource corpus, In Proceedings of the 14th Annual Meeting of the Society for Text and Discourse, 2004

[5] Deza, M. M., Deza E.: Encyklopedia of Distances, Springer, 2009, ISBN: 978-3-642-00233-5

[6] Evert S.: Google Web 1T 5-Grams Made Easy (but not for the computer), In NAACL HLT 2010 Sixth Web as Corpus Workshop, 2010

[7] Fellbaum, Ch.: WordNet: An Electronic Lexical Database, 1998, The MIT Press ISBN-10: 0-262-06197-X, ISBN-13: 978-0-262-06197-1

[8] Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E., Placing Search in Context: The Concept Revisited, *ACM* Transactions on Information Systems, 20(1):116-131, January 2002

[9] Giuliano, C., Gliozzo, A., Gangemi, A., Tymoshenko, K.: Acquiring Thesauri from Wikis by Exploiting Domain Models and Lexical Substitution, In 7th Extended Semantic Web Conference (ESWC2010), 2010

[10] Götze J., Rieder P., Hekstra G. J.: SVD-updating Using Orthonormal μ-Rotations, 1996

[11] Holmes, M. P., Gray, A. G., Isbell, C.L. Jr.: Fast SVD for Large-Scale Matrices, In Workshop on Efficient Machine Learning at NIPS, 2007

[12] Jing Y., Croft B. W.: An Association Thesaurus for Information Retrieval, In RIAO 94 Conference,1994

[13] Kanerva P., Kristoferson J., Holst A. : Random Indexing of Text Samples for Latent Semantic Analysis, In Gleitman, L. R. and Josh, A. K. (Eds.) Proceedings of the 22nd Annual Conference of the Cognitive Science Society (p. 1036), 2000

[14] Krátky, M.: Využití SVD pro indexování latentní sémantiky, Technická zpráva, 2002

[15] Landauer, T. K., Foltz, P. W., Laham, D.: An Introduction to Latent Semantic Analysis, In Discourse Processes, No. 25., pp. 259-284. 1998

[16] Lassi, M.: Automatic thesaurus construction, University Collage of Boras, Sweden, 2002

[17] Lin, D.: An Information-Theoretic Definition of Similarity, In Proceedings of the 15th International Conference on Machine Learning, 1998

[18] Malmkjær, K.: The Linguistics Encyclopedia, 2nd ed, Routledge, 2002, ISBN 0415222109.

[19] Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing, The MIT Press, 1999, ISBN-10: 0-262-13360-1 ISBN-13: 978-0-262-13360-9

[20] Milne, D.: Computing Semantic Relatedness using Wikipedia Link Structure. In Proceedings of NZCSRSC'07, the Fifth New Zealand Computer Science Research Student Conference, 2007

[21] Milne, D., Witten, I. H.: Learning To Link with Wikipedia, In Proceeding CIKM '08 Proceeding of the 17th ACM conference on Information and knowledge management 2008

[22] Moonen, M., von Dooren, M., Vandewalle, J.: An SVD Updating Algorithm for Subspace Tracking, SIAM J. Matrix Anal. Appl., vol. 13, no. 4, pp. 1015–1038,1992

[23] Novák, J.: Sémantická blízkost termínů, bakalářská práce, Brno, FIT VUT v Brně, 2009

[24] Ramos, J.: Using TF-IDF to Determine Word Relevance in Document Queries, In Proceedings of the iCML 2003, 2003

[25] Reynolds, H.T.: The Analysis of Cross-Classifications, New York: The Free Press, 1977

[26] Sahlgren, M.: The distributional Hypothesis. Rivista di Linguistica, volume 20, 2008

[27] Sahlgren, M.: An Introduction to Random Indexing, In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, 2005

[28] Syed, Z., S., Finin, T., Joshi, A.: Wikipedia as an Ontology for Describing Documents, In Proceedings of the Second International Conference on Weblogs and Social Media, AAAI Press, 2008

[29] Tripp, O., Feitelson, D.: Zipf's law Revisited, School of Computer Science and Engineering, The Hebrew University of Jerusalem, Tech. Rep. 2007-115, Aug 2007

[30] Apache Lucene [online] [cit. 2011-4-30] Dostupné na URL: <http://lucene.apache.org/>

[31] nltk.corpus.reader.wordnet-module API [online] [cit. 2010-12-31] Dostupné na URL: <http://nltk.googlecode.com/svn/trunk/doc/api /nltk.corpus.reader.wordnet-module.html>

[32] Get1T [online] [cit. 2011-4-30] Dostupné ne URL: <http://get1t.sourceforge.net/>

[33] semanticvectors [online] [cit 2011-4-30] Dostupné na URL: <http://code.google.com/p/semanticvectors/>

[34] Semantically Annotated Snapshot of the English Wikipedia (SW v.1) [online] [cit. 2010-12-27] Dostupné na URL: < http://barcelona.research.yahoo.net/dokuwiki/doku.php?id= semantically_annotated_snapshot_of_wikipedia>

[35] SVDLIBC. [online] [cit. 2010-12-31] Dostupné na URL: <http://tedlab.mit.edu/~dr/svdlibc/>

[36] SVDPACKC. [online] [cit. 2010-12-31] Dostupné na URL: <http://www.netlib.org/svdpack/>

[37] The Edinburgh Associative Thesaurus [online] [cit. 2010-12-27] Dostupné na URL: <http://www.eat.rl.ac.uk/>

[38] The Karpeles Manuscript Library Museum: Roget's Thesaurus [online] Dostupné na URL: <http://www.rain.org/~karpeles/rogetdis.html>

[39] The WordSimilarity-353 Test Collection [online] [cit. 2010-12-27] Dostupné na URL: <http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/>

[40] thesaurus.com [online] [cit.2011-11-5] Dostupné na URL: <http://thesaurus.com/>

[41] Web 1T 5-gram [online] [cit. 2010-12-31] Dostupné na URL: <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13>

Zoznam príloh

Príloha A. Príklady vygenerovaných podobných termínov pre termíny z WordNetu

V tabuľkách je najpodobnejších 5 termínov pre daný skúmaný termín, ktoré sú podobnejšie ako 0,5. Pokiaľ systém nenašiel 5 podobných termínov, je ich menej a ostatné polia sú prázdne. V prvej tabuľke sú podobné termíny získané metódou Random Indexing a v druhej lexikálnou substitúciou.

Termín	**1**	**2**	**3**	**4**	**5**
cannonballs	volleys	cannon ball	projectile	chariots	bullet
apollo program	apollo	mercury program	gemini program	skylab	nasa
harpoon	exocet	tomahawk	scud	abm	sea wolf
water cannons	water cannon	rubber bullet	batons	tear gas	
paperback book	book	comic book	volume	pamphlet	paper
album	concept album	lps	demo	single	tribute album
novgorod	novosibirsk	leningrad	omsk	murmansk	rostov
wind	drop	turn	winds	expose	drift
iron	old	call	come	rich	stand

Termín	**1**	**2**	**3**	**4**	**5**
cannonballs	shot	bullet	projectile	volleys	chariots
apollo program	apollo	nasa	spacecraft	exploration	space program
harpoon	stinger	tomahawk	nuclear warhead	icbm	scud
water cannons	tear gas	batons	water cannon	rubber bullet	
paperback book	comic book	book	magazine	paperbacks	paper
album	music	song	band	dvd	artist
novgorod	leningrad	novosibirsk	rostov	murmansk	omsk
wind	winds	light	fire	sea	rain
iron	copper	steel	old	wood	small

Príloha B. Užívateľský manuál s popisom činnosti systému

Celý systém na tvorbu tezauru sa spúšťa pomocou skriptu system.sh. V tomto skripte sú spúšťané jednotlivé programy, ktoré tvoria systém. Pred každým príkazom na spustenie programu sú v tomto skripte v komentári uvedené všetky povinné aj nepovinné parametre konkrétneho programu.

V nasledujúcom texte sa ako skúmané termíny chápu tie, ktoré boli systému zadané a pre ktoré má nájsť podobné termíny. Ako kandidátne termíny označujú tie, ktoré boli systému zadané a z ktorých sa vyberajú podobné termíny pre skúmané termíny.

Extrakcia termínov

Najprv sa vykoná extrakcia termínov. Pri nej vzniknú 2 súbory. Jeden bude obsahovať termíny z každého článku Wikipédie a druhý bude obsahovať zoznam termínov.

Príklad prvého súboru:

<DOC Hercule Poirot>
hercule
poirot
hercule poirot
article
…
</DOC>
<DOC Eiffel>
eiffel
eiffel
eiffel
refer
…
</DOC>
…

Druhý súbor obsahuje zoznam termínov z Wikipédie. Každý termín bude na samostatnom riadku. Tento zoznam bude následne filtrovaný.

Indexácia a tvorba kontextových vektorov

Ďalej sa vykoná extrakcia termínov. Pri nej sa použije na vstupe súbor s termínmi z Wikipédie a filtrovaný zoznam termínov. Vznikne index vo formáte Apache Lucene a kontextové vektory uložené v binárnom formáte. Voliteľne je možné spustiť program na vytváranie kontextových

vektorov pre dokumenty (koncepty), ktorý vytvorí kontextové vektory pre dokumenty a uloží ich do rovnakého binárneho formátu ak kontextové vektory pre termíny.

Vyhľadávanie kandidátnych termínov pre skúmané termíny

Kandidátne termíny budú vyhľadávané pomocou porovnávania kontextových vektorov, keď pre každý skúmaný termín sa vyberú najpodobnejšie termíny. Rovnakým spôsobom je možné vyhľadať aj kandidátne termíny pre koncepty. Výstupom je súbor, ktorý obsahuje na jednom riadku skúmaný termín a tabulátormi oddelený zoznam kandidátnych termínov. Voliteľne je možné vypísať aj hodnotu podobnosti, ktorá bude tiež oddelená tabulátorom. Slová vo viacslovných termínoch sú oddelené medzerou.

Príklad súboru (s vypísanou hodnotou podobnosti):

loosestrife purple loosestrife 0.5683233
admission fee entrance fee 0.57309544 admission charge 0.54373866
television monitor television set 0.5087565

Príklad bez hodnoty podobnosti:

loosestrife purple loosestrife
admission fee entrance fee admission charge
television monitor television set

Súbor s kandidátnymi termínmi pre koncepty má rovnakú štruktúru. Pri ďalšom spracovaní sa používa súbor bez vypísaných hodnôt. Systém je však schopný korektne spracovať aj súbor s vypísanými hodnotami, ale číselné hodnoty podobnosti bude brať ako ďalšie termíny. Pre ne sa však nenájdu zodpovedajúce súbory s kontextami, takže budú mať nulovú podobnosť a budú zaradené na koniec zoznamu podobných termínov.

Vyhľadávanie kontextov pre skúmané termíny

Vyhľadávanie prebieha v SW1 korpuse. Nájdené kontexty sú najprv uložené do niekoľkých súborov (mená sú čísla) do určeného adresára. Každý súbor obsahuje všetky kontexty pre niekoľko termínov. Na jednom riadku je vždy termín a kontext oddelený tabulátorom. Výskyt termínu v kontexte je nahradený za žolíkový znak "<*>"

Príklad obsahu súboru (0.tmp):

hallmark <*> of the gangster
hallmark <*> of the gangster period
impress <*> some people
impress <*> by nash 's vocal

Do samostatných súborov sú uložené aj ľavé a pravé kontexty (ktoré sú pred a za termínom). Uložené sú podľa počtu slov v takomto kontexte. Každý súbor obsahuje maximálne množstvo pravých a

ľavých kontextov, ktoré je zadané parametrom. Tieto ľavé a pravé kontexty sa potom vyhľadávajú vo Web 1T korpuse pomocou nástroja Get1T.

Súbory s nájdenými kontextami pre skúmané termíny sú následne prevedené z tohto medzikroku do konečného formátu. Ten vyzerá tak, že pre každý skúmaný termín je jeden súbor, ktorý obsahuje kontexty pre tento termín. Názov súboru je zhodný s termínom. Súbory sú uložené v zadanom adresári v podzložkách, ktoré sú nazvané podľa prvých dvoch znakoch termínov, ktoré sú uložené v danom podadresári.

Príklad obsahu súborov (súbor affair, uložený v podzlože af):

<*> specialty
<*> analysis
foreign <*>
klos airs a public <*>
public <*>
airs a public <*>

Vyhľadávanie kontextov pre kandidátne termíny

Zo zoznamu zadaných kandidátnych termínov sa vytvoria všetky ich všetky možné kontexty pomocou žolíkových znakov "<*>". Tieto kontexty sa následne vyhľadajú pomocou nástroja Get1T. Kontexty budú rozdelené podľa dĺžky

Príklad súboru s kontextami dĺžky 5:

product line <*> <*> <*>
<*> product line <*> <*>
<*> <*> product line <*>
<*> <*> <*> product line
hadrosaurus <*> <*> <*> <*>
<*> hadrosaurus <*> <*> <*>
<*> <*> hadrosaurus <*> <*>
<*> <*> <*> hadrosaurus <*>
<*> <*> <*> <*> hadrosaurus

Filtrovanie a ohodnotenie kontextov pre kandidátne termíny

Nájdené kontexty pre kandidátne termíny budú filtrované pomocou pravých a ľavých kontextov nájdených pre skúmané termíny. Najprv sa prevedú zo súboru, ktorý obsahuje len zoznam n-gramov do medzikroku. To sa spraví tak, že určí ktorý nájdený n-gram obsahuje aký termín (alebo termíny). Následne sa n-gram rozdelí na termín, ľavý a pravý kontext a frekvenciu výskytu. Takto získané kontexty sa zapíšu do súboru do zadaného adresára. Každý súbor obsahuje kontexty pre niekoľko termínov. Na každom riadku je termín, ľavý a pravý kontext, dĺžka ľavého a pravého kontextu, celková dĺžka n-gramu (ľavý + pravý kontext + termín) a frekvencia výskytu n-gramu. Za dĺžku

sa považuje počet slov. Jednotlivé stĺpce sú oddelené tabulátorom. Pokiaľ je dĺžka niektorej časti nulová, v súbore sú za sebou na tomto mieste 2 tabulátory.

Príklad obsahu súboru (0.tmp):

times ä°spanyol mastiff visited 53 4 0 5 1037
manufacturing Í edexcel gcse in 4 0 5 79
times Ífall seven , stand 2 2 5 104
potter Íharry and the goblet 1 3 5 80
times Í″ june 7 : 4 0 5 46
times Í″ sept 10 : 4 0 5 46

Následne sú tieto kontexty filtrované. Vyhadzujú sa tie, ktoré nemajú pravý alebo ľavý kontext medzi pravými a ľavými kontextami, ktoré boli nájdené pri vyhľadávaní kontextov pre skúmané termíny. Tie kontexty, ktoré prejdú filtrom sa ohodnotia ako podiel ich PMI a SI a sú uložené do výsledného formátu. Ten vyzerá tak, že pre každý skúmaný termín je jeden súbor, ktorý obsahuje kontexty pre tento termín. Tento súbor obsahuje na každom riadku kontext a jeho ohodnotenie oddelené tabulátorom. Názov súboru je zhodný s termínom. Súbory sú uložené v zadanom adresári v podzložkách, ktoré sú nazvané podľa prvých dvoch znakoch termínov, ktoré sú uložené v danom podadresári.

Príklad obsahu súboru (súbor buena vista z podzložky bu):

back by <*> home 1.2616660028849802
balboa park , <*> 1.7598967857408516
baptist church in <*> 1.8489106045694177
bar at <*> international 1.3205885272202154
battle of <*> battle 1.3253716455907325
bay <*> buena vista 1.6026377769699345
beach , <*> beach 1.3517663888063185
beach , <*> county 1.3895509037623313

Ohodnotenie termínov pomocou lexikálnej substitúcie

Ohodnotenie kandidátov pomocou lexikálnej substitúcie prebieha porovnávaním ich kontextov. Na vstupe je pre každý skúmaný termín zoznam kandidátnych termínov a kontexty pre skúmané a kandidátne termíny. Tento zoznam vznikol v kroku vyhľadávania kandidátnych termínov pre jednotlivé skúmané termíny. V kroku lexikálnej substitúcie je tento zoznam pre každý termín preusporiadaný na základe výsledkov lexikálnej substitúcie. Formát výstupu zostáva rovnaký ako vstupný.

Príloha C. CD so zdrojovými textami, programovou dokumentáciou a ukážkami výsledkov

Na priloženom CD sa nachádzajú nasledujúce adresáre:

- System - obsahuje funkčnú verziu systému na automatickú tvorbu tezauru z Wikipédie. Obsah toho adresára je:
  - get1t-0.2.2 – adresár, ktorý obsahuje zdrojové kódy nástroja Get1T
  - src – adresár, ktorý obsahuje zdrojové kódy častí systému napísaných v Jave
  - Thesaurus.jar – jar súbor s časťami systému napísanými v jave
  - semanticvectors-1.26.jar – knižnica semanticvectors
  - lucene-core-2.3.2.jar – knižnica Apache Lucene
  - system.sh – skript, ktorý spúšťa jednotlivé súčasti systému v správnom poradí
  - README – súbor, ktorý obsahuje popis činnosti systému
  - stoplist.txt – súbor, ktorý obsahuje stoplist pre filtrovanie termínov
- Dokumentacia – programová dokumentácia vygenerovaná nástrojom javadoc
- Vysledky – ukážky vygenerovaných výsledkov
- technicka_sprava.pdf – technická správa vo formáte PDF